University of Alberta

Systems Approach to Managing Educational Quality in the Engineering Classroom

by

Kostyantyn Grygoryev ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the

requirements for the degree of *Doctor of Philosophy*

Department of *Mechanical Engineering*

Edmonton, Alberta
Fall 2005

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

0-494-08647-5

Your file Votre référence
ISBN:
Our file Notre reference
ISBN:

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

ABSTRACT

Today's competitive environment in post-secondary education requires universities to demonstrate the quality of their programs in order to attract financing, and student and academic talent. Despite significant efforts devoted to improving the quality of higher education, systematic, continuous performance measurement and management still have not reached the level where educational outputs and outcomes are actually produced – the classroom.

An engineering classroom is a complex environment in which educational inputs are transformed by educational processes into educational outputs and outcomes. By treating a classroom as a system, one can apply tools such as Structural Equation Modeling, Statistical Process Control, and System Dynamics in order to discover cause-and-effect relationships among the classroom variables, control the classroom processes, and evaluate the effect of changes to the course organization, content, and delivery, on educational processes and outcomes.

Quality improvement is best achieved through the continuous, systematic application of efforts and resources. Improving classroom processes and outcomes is an iterative process that starts with identifying opportunities for improvement, designing the action plan, implementing the changes, and evaluating their effects. Once the desired objectives are achieved, the quality improvement cycle may start again.

The goal of this research was to improve the educational processes and outcomes in an undergraduate engineering management course taught at the University of Alberta. The author was involved with the course, first, as a teaching assistant, and, then, as a primary instructor. The data collected from the course over four years were used to create, first, a static and, then, a dynamic model of a classroom system. By using model output and qualitative feedback from students, changes to the course organization and content were introduced. These changes led to a lower perceived course workload and increased the students' satisfaction with the instructor, but the students' overall satisfaction with the course did not change significantly, and their attitude toward the course subject actually became more negative.

This research brought performance measurement to the level of a classroom, created a dynamic model of the classroom system based on the cause-and-effect relationships discovered by using statistical analysis, and used a systematic, continuous improvement approach to modify the course in order to improve selected educational processes and outcomes.

TABLE OF CONTENT

Chapter 1: Introduction 1
1.1. Performance measurement and education 1
1.2. Monitoring educational processes 3
1.3. Discovering cause-and-effect relationships 4
1.4. Designing and testing policies 5
1.5. Continuous improvement 6
1.6. Organization of the thesis 7
Chapter 2: Literature Review 8
2.1. Introduction 8
2.2. Performance measurement 8
    2.2.1. Performance measurement in everyday life 8
    2.2.2. Performance measurement in the public sector 9
    2.2.3. Performance measurement in education 10
    2.2.4. Organizational performance measurement frameworks in education 12
    2.2.5. The Balanced Scorecard in post-secondary education 13
    2.2.6. Measuring educational quality at the classroom level. 14
    2.2.7. Additional use of student survey data 16
2.3. Monitoring educational processes 16
    2.3.1. Pursuit of quality 16
    2.3.2. Statistical Quality Control tools in education 18
2.4. Discovering cause-and-effect relationships 20
    2.4.1. Limitations of ordinary regression models 21
    2.4.2. From ordinary regression to Structural Equation Modeling 23
    2.4.3. SEM in Education 25
    2.4.4. Limitations of SEM models 27
        2.4.4.1. Assessment of fit versus assessment of causal claims 28
        2.4.4.2. Effect of time 29
        2.4.4.3. Time-dependent inputs and feedback loops 30
2.5. Designing and testing policies by using System Dynamics 31

2.5.1. What is System Dynamics? 31
2.5.2. System Dynamics in education 33
2.6. Summary 35
Chapter 3: Motivation, Objectives, and Methodology 36
3.1. Introduction 36
3.2. Motivation 36
3.2.1. Performance Measurement 36
3.2.2. Process Control 38
3.2.3. System Modeling 39
3.2.4. System Analysis 42
3.3. Objectives 43
3.4. Methodology 44
3.5. Performance framework 46
3.5.1. Classroom as a system 46
3.5.2. Inputs, outcomes, and their indicators 48
3.5.3. The survey 49
3.5.4. Educational processes 50
3.5.5. Modified Background Knowledge Probe – a tool to measure knowledge transfer 52
3.6. Limitations 55
3.7. Summary 55
Chapter 4: Control of Educational Processes 56
4.1. Introduction 56
4.2. Implementation 56
4.3. Statistics computed from an MBKP 58
4.4. Use of the collected data 59
4.5. Example 60
4.5.1. Data source 60
4.5.2. Statistical principles 62
4.5.3. SPC charts 64
4.5.4. Discussion of results: Instructors A and B (Fall 2002) 65

4.6.5. Discussion of results: Instructor A (Winter and Fall 2002) 69
4.5.6. Discussion of results: Instructor A (Fall 2002 and Winter 2003) 71
4.5.7. Autocorrelation considerations 74
4.5.8. Warm-up instability 76
4.5.9. Short-run process 78
4.6. Attributes vs. variables SPC chart: the problem of a small class 81
4.7. MBKP with the statistic as variable: application 83
4.8. Search for assignable causes 89
4.9. Suggestions for the Design of "Before and After" Questions 92
4.10. A designed experiment as a tool to discover assignable causes 93
4.11. Limitation 94
4.12. Summary 94
Chapter 5: Discovering Cause-And-Effect Relationships 96
5.1. Introduction 96
5.2. Model variables and hypothesized relationships 96
5.3. Data 102
5.4. LISREL model 105
5.5. Model details 109
5.5.1. Fixed λ coefficients 109
5.5.2. Fixed indicators' measurement errors 110
5.5.3. Multiple indicators 114
5.5.4. Reciprocal effects 116
5.5.5. Loops 118
5.5.6. Control variables 119
5.6. Model results and assessment of fit 120
5.6.1. Looking beyond fit indices 122
5.6.2. Squared Multiple Correlations 122
5.6.3. Multiple indicators 126
5.6.4. Analysis of residuals 127
5.6.5. Reciprocal effects and correlation among the estimates 128
5.7. Model Improvement 129

5.7.1. Adjustments to the SEM model: non-significant coefficients 130

5.7.2. Adjustments to the SEM model: modification indices 131

5.8. Adjusted model: results and discussion 133

5.9. Summary 136

Chapter 6: Design and Testing of Policies by Using System Dynamics 137

6.1. Introduction 137

6.2. System Dynamics modeling steps 137

6.3. SD Model 138

6.3.1. Problem definition 138

6.3.2. Development of a dynamic hypothesis 139

6.3.2.1. Balancing loop "Studying" 140

6.3.2.2. Balancing loops "Workload effect I" and "Workload effect II" 141

6.3.2.3. Reinforcing loops "Attitude effect I" and "Attitude effect II" 142

6.3.2.4. Complete causal model 143

6.3.3. Model construction 144

6.3.3.1. Stock-and-flow map based on the loop "Studying" 144

6.3.3.2. Model equations and starting values 146

6.3.3.3. Model behaviour 148

6.3.3.4. Model testing 149

6.3.4. Complete STELLA model 151

6.3.5. Model equations and starting values 152

6.3.6. Behaviour of the complete model 153

6.3.7. Model testing 157

6.3.7.1. Satisfaction with the course and the instructor, and attitude toward subject 157

6.3.7.2. Sensitivity analysis: the "good" versus the "bad" student 160

6.3.7.3. Extreme values 163

6.3.7.4. Variables' ranges and scales 165

6.3.8. Policy testing 166

6.3.8.1. Assignment policy 167

6.3.8.2. Attendance policy 170

6.4. Summary 174

Chapter 7: Course Improvement 175

7.1. Introduction 175

7.2. Continuous improvement 175

7.3. Application 176

  7.3.1. Plan 176

  7.3.2. Do 178

  7.3.3. Check 179

    7.3.3.1. After-midterm survey 179

      7.3.3.1.1. Comparison between Winter 2005 and Fall 2003 – Winter 2004 182

      7.3.3.1.2. Comparison between Sections X and Y, Winter 2005 182

      7.3.3.1.3. Instructor Y, Winter 2004 and Winter 2005 184

    7.3.3.2. SEM model 185

    7.3.3.3. Knowledge transfer 187

  7.3.4. Act 189

7.4. Afterword 192

7.5. Summary 192

Chapter 8: Conclusions 194

8.1. Contributions of the research 194

8.2. Directions of future research 196

Bibliography 199

Appendices 213

Appendix I: Performance Measurement in the Public Sector 214

Appendix II: Modeling Theory 227

Appendix III: Structural Equation Modeling Basics 232

Appendix IV: Systems Thinking 239

Appendix V: Application for Study Approval 248

Appendix VI: Before and After questions appearing in MBKP sets 1, 2, and 3 251

Appendix VI: Before and After questions appearing in MBKP sets 1, 2, and 3 252

Appendix VII: Data Collected From MBKPs 259

Appendix VIII: Overview of Statistical Quality Control 261
Appendix IX: Before and After Questions Appearing in "Variable" MBKPs 265
Appendix X: Variable Knowledge Transfer Statistic, Raw Data 267
Appendix XI: Use of Designed Experiments to Discover Assignable Causes 270
Appendix XII: Matrix of Covariances Among Observed Indicators (Matrix S), SEM Classroom Educational Model 274
Appendix XIII: SEM Models "Original" and "Final" in a Path-Diagram Form 275
Appendix XIV: LISREL Syntax, Model "Original" 275
Appendix XIV: LISREL Syntax, Model "Original" 276
Appendix XV: System Dynamics Modeling Steps 278
Appendix XVI: SD Model Equations 287
Appendix XVII: Before and After Questions Appearing in MBKPs Administered in Winter 2005 289
Appendix XVIII: Statistical Analysis of Survey Data, Winter 2005 294
Appendix XIX: SEM Model "Winter 05" in a Path-Diagram Form 301
Appendix XX: *S* matrix, Winter 2005 302
Appendix XXI: LISREL Syntax, Model "Winter 05" 303
Appendix XXII: Data Collected From MBKPs, Winter 2005 305

LIST OF TABLES

Table 3.1: Student inputs, outcomes, and background characteristics 51

Table 4.1: B&A answer combinations... 58

Table 4.2: Description of MBKP data sets 61

Table 4.3: MBKP data sets used in the analysis 62

Table 4.4: Absolute differences in performance (Instructors A and B, Fall 2002) 68

Table 4.5: Instructor A, absolute differences in performance (Winter and Fall 2002)..... 70

Table 4.6: Instructor A, absolute differences in performance (Fall 2002 and Winter 2003) 74

Table 4.7: Results of the Durbin-Watson autocorrelation tests 76

Table 4.8: Data summary, variable statistic.. 85

Table 5.1: Variables and indicators in the SEM model of classroom performance 98

Table 5.2: Postulated causal relationships in the SEM model of classroom performance 99

Table 5.3: Student questionnaire and corresponding SEM variables 103

Table 5.4: Proportion of indicators' fixed error variance 111

Table 5.5: LISREL estimates of the structural coefficients β and γ, model "Original".. 121

Table 5.6: Squared multiple correlations for latent endogenous variables..................... 124

Table 5.7: Structural effects removed from the SEM model 131

Table 5.8: Statistically significant modification indices.. 132

Table 5.9: Structural coefficients β and γ values for the original and adjusted models . 135

Table 6.1: Parameters for test score sensitivity analysis scenarios................................. 158

Table 6.2: "Good" versus "bad" student model parameters 161

Table 6.3: Assignment policies... 168

Table 7.1: Course improvement plan.. 177

Table 7.2: Absolute difference in performance, 2002-2003 and 2005 189

Table XI.1: Results of completely randomized single-factor design (adopted from Montgomery 1997b) 271

Table XI.2: ANOVA table, completely randomized single-factor design (adopted from Montgomery 1997b).. 272

Table XI.3: Factorial design for testing "poor lecture and poor question" hypothesis... 273

Table XV.1: SD model tests of structure (based on Sterman 2000, and Forrester and Senge 1980) 283

Table XV.2: SD model tests of behaviour (based on Sterman 2000, and Forrester and Senge 1980) 285

Table XV.3: SD model tests of policy implications (based on Forrester and Senge 1980) 286

Table XVIII.1: Goodness-of-fit test, midterm marks, section X, Winter 2005 294

Table XVIII.2: Goodness-of-fit test, midterm marks, section Y, Winter 2005 294

Table XVIII.3: Comparison of survey results between Winter 2005 and 2003-2004 295

Table XVIII.4: Comparison between sections X and Y, Winter 2005 semester 297

Table XVIII.5: Comparison between instructor Y's results, Winter 2004 and Winter 2005 299

LIST OF FIGURES

Figure 3.1: Research methodology 44

Figure 4.1: Algorithm for the design and administration of MBKPs 57

Figure 4.2: P_{IBCA} standardized attributes chart, Instructors A and B, Fall 2002 66

Figure 4.3: P_{IA} standardized attributes chart, Instructors A and B, Fall 2002 67

Figure 4.4: P_{IBCA} standardized attributes chart, Instructor A, Winter and Fall 2002 70

Figure 4.5: P_{IA} standardized attributes chart, Instructor A, Winter and Fall 2002 71

Figure 4.6: P_{IBCA} standardized attributes chart, Instructor A, Fall 2002 and Winter 2003 72

Figure 4.7: P_{IA} standardized attributes chart, Instructor A, Fall 2002 and Spring 2003 ... 72

Figure 4.8: P_{IBCA} standardized attributes chart, first four points discarded, Fall 2002 78

Figure 4.9: P_{IBCA} Q chart, Fall 2002 80

Figure 4.10: Example of a MBKP collecting a variable statistic 84

Figure 4.11: Sample mean SPC chart, variable characteristic 86

Figure 4.12: Sample standard deviation SPC chart, variable characteristic 86

Figure 4.13: Algorithm for searching for an assignable cause 91

Figure 5.1: SEM model with fixed indicators' error variances 113

Figure 5.2: Latent variable "Time devoted to self-studying" with multiple indicators .. 114

Figure 5.3: Reciprocal effect between latent variables in the SEM model 117

Figure 5.4: Loop "Workload Effect I" 118

Figure 5.5: LISREL estimates for variable η_4 with multiple indicators 126

Figure 6.1: Balancing loop "Studying" 140

Figure 6.2: Balancing loops "Workload effect I" and "Workload effect II" 141

Figure 6.3: Reinforcing loops "Attitude effect I" and "Attitude effect II" 142

Figure 6.4: Complete causal SD model 143

Figure 6.5: STELLA stock-and-flow model based on the loop "Studying" 146

Figure 6.6: Behaviour of the variables "Test score" and "homework time" in the model based on the loop "Studying" 148

Figure 6.7: Behaviour of "Test score" under extreme values of the assignment rate 150

Figure 6.8: Balance state of the model based on the loop "Studying" 151

Figure 6.9: Stock-and-flow map for the complete STELLA model 152
Figure 6.10: Behaviour of the variables "Test score" and "Homework time" in the complete model 153
Figure 6.11: Behaviour of the variables "Satisfaction with the course," "Attitude toward the subject," and "Perceived workload" in the complete model.................. 154
Figure 6.12: Shift in loop dominance 155
Figure 6.13: Behaviour of the complete model over a 40-week time horizon 156
Figure 6.14: Behaviour of the complete model over a 70-week time horizon 157
Figure 6.15: Sensitivity of the variable "Test score" to the changes in the variable "Satisfaction with the course" 158
Figure 6.16: Sensitivity of the variable "Test score" to the changes in the variable "Satisfaction with the instructor" 159
Figure 6.17: Sensitivity of the variable "Test score" to the changes in the variable "Attitude toward the subject" 159
Figure 6.18: Behaviour of the variable "Test score" under the various student scenarios 161
Figure 6.19: Peak homework time 162
Figure 6.20: "Good" versus "bad" student: timing of increase in homework time 163
Figure 6.21: Behaviour of the variable "Test score" under the zero homework hours scenario 164
Figure 6.22: Behaviour of the variable "Test score" in the complete model under the extreme values of the variable "Assignment rate" 165
Figure 6.23: Behaviour of the variable "Test score" under various assignment policies 168
Figure 6.24: Behaviour of the variable "Attitude toward the subject" under various assignment policies 169
Figure 6.25: Behaviour of the variable "Test score" under various attendance scenarios 171
Figure 6.26: Rise in absence rate, scenario #5 172
Figure 6.27: Behaviour of the variable "Attitude toward the subject" under various attendance scenarios 173
Figure 7.1: Midterm marks distribution, true and reported, section X 181

Figure 7.2: Midterm marks distribution, true and reported, section Y 181
Figure 7.3: Knowledge transfer SPC chart, Instructor X, Winter 2005........................ 188
Figure III.1: SEM model in a pictorial form... 234
Figure IV.1: Systems Thinking theoretical approaches.. 244
Figure VIII.1: Example of a statistical control chart 263

Chapter 1: Introduction

The importance of education to a nation's economic development, security, and scientific progress cannot be underestimated. In today's economic environment, knowledge is the most important source of competitive advantage. Governments around the world recognize this importance and invest considerable resources in educational systems. In return for the resources received, educational institutions have to demonstrate the outcomes and outputs produced. Both the general public and educational professionals want to be sure that taxpayer dollars are well spent, and that the students being educated today are well prepared to become the productive citizens of tomorrow. The government, in turn, has a responsibility to spend taxpayers' money efficiently (McLellan 2005).

Post-secondary educational institutions enjoy a greater degree of autonomy from the government's formal administrative offices than educators at the elementary school level. The professional independence of faculty reveals this autonomy, and some educational organizations enjoy it to a greater extent than the others, but it comes at the price of increased expectations and the increased accountability educators have to provide for their outcomes.

1.1. Performance measurement and education

The private sector embraced comprehensive performance measurement when organizations realized that traditional financial measures no longer revealed the true state of organizational health. The nature of the economy has changed from managing tangible assets such as equipment, inventory, and buildings, to managing intangible assets such as information, databases, customer and supplier relations, and management knowledge and practices. These new assets became a source of competitive advantage, but without proper measurement tools organizations encountered difficulties in managing them (Andersen and Fagerhaug 2002).

Eventually, performance measurement practices found their way into the non-profit and public sectors as well. With the ever-increasing cost of post-secondary education, universities have to demonstrate high-quality education to attract students, academics, and government and private funding. Universities try to ensure the quality of their educational outcomes and outputs through various methods, such as accrediting faculties and departments, measuring students' achievement and satisfaction with instruction, and measuring the quality of academic work by the number of articles published and research grants received by faculties. In some countries, universities even register their educational programs to the ISO 9001 standard (Karapetrovic 1998).

While much research has been conducted on improving the post-secondary educational system, educators at the classroom level have not yet widely adopted performance measurement. The lack of well-developed tools and methods for measuring educational processes and outcomes leaves many questions unanswered. Educators in a classroom, where quality education is actually being produced, are still looking for answers to questions:

- What processes are responsible for producing better knowledge and better students?
- What inputs, processes, and outputs should be measured?
- How can an educational system be modeled?
- What effect will my decisions have on the outcomes?

This thesis will attempt to develop a model for the systematic measuring of teaching and learning effectiveness. The model will be used to re-design an undergraduate engineering management course with the goal of improving student performance, satisfaction with the instructor and with the course, and the students' attitude toward the subject of the course.

The current developments in the fields of monitoring, modeling, and managing educational performance will be assessed by reviewing academic research, institutional policies, and legislative initiatives. The literature review is presented in Chapter 2. Chapter 3 will explain how problems identified in the literature review provided motivation for this work. Chapter 3 also presents a methodology for a systematic approach to controlling and managing educational performance in a classroom.

1.2. Monitoring educational processes

At the post-secondary level, systematic experimental research on teaching and learning began during the 1930s. Researchers started analyzing educational processes, in addition to their resulting products, in the 1980s (McKeachie 1990). However, educational processes are still widely viewed as "black boxes," while process inputs and outputs remain the main focus of study (Tam 2001).

In post-secondary engineering education, the students' interaction with the instructor in a classroom is one of the main sources of learning. Nevertheless, very few tools are available to quantitatively evaluate, in a systematic way, the instructor's teaching effectiveness in transferring knowledge to students. Measuring students' performance on homework assignments and midterm and final exams is not sufficient for two reasons. First, midterm and final exam data become available too late in the semester, when only a retrospective analysis can be conducted to discover causes of poor or good performance. Second, the instructor may be interested in her/his own teaching effectiveness. Student performance on a test is an indicator of a student's knowledge gain from the classroom interaction with the instructor and, perhaps even more importantly, the knowledge gained from "self-education" and peer learning. Research indicates that independent and peer study can be as effective as the professor-led instruction (Jenkins 2003, McKeachie 1990).

Industry has long ago realized that in order to provide a quality product or service, controlling the processes responsible for the production of the output is cheaper and more effective to, than inspecting the output itself. Control at the end of the process, even when conducted in a rigorous and scientific way, is "too late, costly, and ineffective," as Dr. W. Edwards Deming's stated (as cited in Woodhall and Montgomery 1999, p. 376). Chapter 4 of this thesis presents a method for the continuous monitoring of knowledge transfer and describes how statistical quality control tools, such as a control chart, can be applied in a university classroom to evaluate the effectiveness of the process of the transfer of knowledge from an instructor to students.

1.3. Discovering cause-and-effect relationships

To determine what factors are responsible for student performance and attitude toward a subject, one must establish the nature of the cause-and-effect relationships among the variables in a classroom educational system. Since the exact nature of the relationship between variables such as, for example, perceived course workload and satisfaction with the instructor, cannot be established, an empirical model of a classroom system must be created.

Regression models, which are frequently used to model phenomena in the field of education, have limitations, which sometimes either are not realized or are ignored by researchers. In a regression model, the independent variables are assumed to be fixed at levels pre-determined by the researcher, are assumed to be measured with a negligible error, and are assumed not to be a subject to random variation. The only variable subject to a random error is the dependent variable (Montgomery and Runger 1999). However, when observation, and not experimentation, is the only way to collect data (as is often the case in social systems (Hox and Bechger 1998, Jöreskog and Sörbom 1996a), all variables, independent and dependent, are subject to a random error and uncontrolled variation (Jöreskog and Sörbom 1996a). Additionally, concepts of interest, such as attitude, cannot be observed directly, and indicators of such concepts are often biased and subject to a significant measurement error (Hayduk 1996, Hayduk 1987, Bartholomew 1983).

Chapter 5 concentrates on creating a model of student performance in an undergraduate engineering class, taking into consideration both the observational character of the data collected in the classroom, and the potential biases introduced into measurements by specific data-collection methods.

1.4. Designing and testing policies

The purpose of modeling is two-fold: to gain insight into a problem and to solve it (Sterman 2000). Researchers gain insight into a problem by postulating causal hypotheses about the relationships among the variables in the system under study and by creating and testing a model of the system. To solve the problem the system and policy changes that will produce the desired outcome must be determined. Can regression models be used to solve the problems of improving the quality of educational processes and outcomes at the level of a classroom?

One of the main limitations of regression-based models is the absence of time as a variable in the model specifications and analysis. Time, though, is a part of the real causal world. In statistical modeling, even when issue of time is raised, a typical approach is to think of the effect as occurring instantaneously (Hayduk 1987), an approach which is a forced approximation. Static analysis implies that the data set on hand was obtained by maintaining inputs at a constant level for a sufficient time for the system to settle into its output values. Even when inputs are controlled at steady levels, if outputs are measured too early, the change in the dependent variables might not have occurred yet. Therefore, one must wait until the system settles into its new values (Heise 1975).

The presence of time-dependent inputs and feedback structures in a system further complicates model analysis. Feedback produces a dynamic non-linear behavior in the system variables, and the variables involved in the feedback relationship arrive at the final values in a non-linear manner (Forrester 1968). While regression analysis can be used to estimate the direct, indirect, and total effects of the variables in systems containing feedback loops (see, for example, Hayduk 1996, Hayduk 1987), the processes by which a system arrives at those values could be of equal or even of greater interest. In fact, this non-linear behavior might, at times, destroy the system before it ever achieves its final static values (Heise 1975). Analysis of the time-dependent behavior of a classroom educational model containing feedback structures and the effects of policy changes on the students' performances and attitudes toward the course subject are presented in Chapter 6.

1.5. Continuous improvement

Every product or service should be designed and produced with a customer in mind. While defining the "customer" of the educational system is a difficult task, at the classroom level, the student is the obvious candidate. If the students' knowledge and satisfaction are the products of one semester of engineering education, then the instructor's role in the classroom should be to improve these two parameters.

Quality improvements are best achieved through the systematic and continuous application of effort. Chapter 7 will demonstrate how the continuous improvement philosophy can be applied to achieve desired outcomes such as enhancing the students' satisfaction with the course and the instructor, and improving the students' attitudes toward the subject.

1.6. Organization of the thesis

The remainder of the thesis is organized in the following way. Chapter 2 presents a review of literature on performance measuring, system modeling and analysis, and policy testing and design, as applied in post-secondary education. The motivation for and the objectives of the research are presented in Chapter 3, together with an outline of the overall structure of the approach used to improve student performance and satisfaction in the classroom. Chapter 4 proposes a tool, based on the Statistical Process Control (SPC) methodology, to measure and monitor the process of the knowledge transfer between the instructor and the students. Chapter 5 addresses the issue of modeling systems in which the data were obtained through observation via application of the Structural Equation Modeling (SEM). Chapter 6 examines the time-dependent behavior of the variables of the classroom system in the presence of feedback structures and analyzes the effect of changes to an undergraduate engineering management course, in terms of the students' performance and attitudes toward the subject, by using the System Dynamics (SD) modeling approach. Chapter 7 describes how changes to the course were introduced based on the insights gained from the system modeling and analysis, and the effect those changes had on students' satisfaction and performance. Finally, the contributions of this research and directions for future research are given in Chapter 8.

Chapter 2: Literature Review

2.1. Introduction

This chapter presents a literature review of research on performance measurement, statistical process control, structural equation modeling, system dynamics, and the application of each of these tools in the field of post-secondary education.

2.2. Performance measurement

2.2.1. Performance measurement in everyday life

While we may not think about performance measurement, we measure performance continually as we go about our daily routine. A student wakes up in the morning, takes a bus to university, attends lectures and laboratory sessions, at the end of the day goes to the stadium to watch a hockey game, returns home, and goes to bed. Let us consider what would happen if a student did not measure his or her own performance. Without checking the alarm clock, he might get up late, miss a bus and be late for classes. At university, if too much time and effort is spent on one particular course (out of, say, six), the mark in that course could be really high, but the student's performance in other courses might suffer due to the lack of time devoted to them. Moreover, if the student's favorite hockey team's quality of the play declines, he will stop going to their games, and the team will lose revenue.

Under closer examination, we realize that the student's routine has all the elements of the performance measurement process that both profit and non-profit organizations use. Performance measurement is an assessment of an organization's performance that has, as its components, measures of productivity, effectiveness, quality, and timeliness (GAO 1980). In the public sector it would be the process of determining how effectively and efficiently taxpayer resources are being used for the delivery of services and the administration of programs (Foltin 1999).

Companies measure performance to remain competitive, i.e., to stay in business. Governments measure performance to overcome budget deficits, reduce taxes, and to improve the quality of life of their citizens. Individuals measure performance to stay in good physical shape and to maintain healthy relationships, or, in other words, to improve their quality of life.

2.2.2. Performance measurement in the public sector

Performance measurement originated in the business sector, but, eventually, was adopted by the public and non-profit organizations. In some instances, it was a "firefighting" measure introduced in an atmosphere of crisis, when government and organization budgets year after year operated in deficit. In other instances, even without any obvious financial crisis, the managers of public organizations realized that aging infrastructure, inflation, and the resistance of constituents to bearing higher costs required improvements in operational efficiency. Public discontent with the quality of provided services forced the public sector to introduce reforms and create a system for public accountability (Foltin 1999, Ogata and Goodkey 1998; Hillison et al. 1995). As foreign competition revealed the need for quality improvement in the business sector, the same quality drive was introduced into the public sector, with the taxpayer viewed as the "customer" (Foltin 1999, Glover 1992).

Appendix I presents an overview of performance measurement in the public sector: the theoretical foundation, criteria for selecting performance indicators, stages of developing a performance measurement system, obstacles to the implementation of performance measurement, and organizational performance frameworks.

2.2.3. Performance measurement in education

The need for high-quality education is recognized at all levels of society and the highest levels of government. United States President George W. Bush stated "We need an educated workforce to keep this country the most productive in the world" (US Budget 2005).

Debate on the quality versus the cost of education takes place at both elementary (Willmore 2004) and post-secondary (Rauf 2004) levels; on college campuses (Karbani 2005, Cairney 2002), in regional (Flower 2004, Conley and Picus 2003) and national (Wells 2004) governments, and around the world (Kripalani 2004). Measuring the cost of education is relatively straightforward – for example, one can do so by analyzing one's tuition bill or the government's budget figures. For example, the U.S. Department of Education has discretionary budget authority of US$57.3 billion for the fiscal year 2005 (US Budget 2005). In Canada, in 2003-2004, the consolidated federal, provincial, territorial and local government expenditures on education were CAD$68.5 billion, or 14.8% of all government expenditures (Statistics Canada 2005).

Measuring the quality of education, however, is a more complicated issue. For a product (or service) to be considered of "high quality", it has to satisfy the requirements of its users (Montgomery 1997a). In education, as in any other non-profit sector, the number of stakeholders is much higher than in the business sector. For a business, major stakeholders are the management and employees (the company itself), as well as the shareholders. Customers, if a company is not a monopoly, can always turn to a competitor. Therefore, the customers are not really concerned about the company's well-being (i.e., if Toyota goes bankrupt, I can always buy a Honda). Therefore, the small number of business stakeholders makes the selection of performance indicators easier – shareholders will track the stock price, and management and employees will concentrate on net income and net cash flow.

A public sector organization, such as a university, is in a much more difficult position. An alternative provider of the government's service or product may not be available, or the alternatives may not be affordable for a majority of populations. Defining the "customer" of a public organization is not as straightforward as in private sector. In the private sector, a person who pays for a service or product is usually the same person who receives a service or product, whereas for a public organization, not only the user of a service or product pays for it (Kaplan 2001, Kaplan and Norton 2001, Smith 1995). In post-secondary education, students pay a portion of the overall costs, and donors or the taxpayers provide the rest. For example, in 2003 in the province of Alberta, Canada, tuition provided 28% of the operating revenues for universities (an increase from 8% in 1983 and 16% in 1993) (CAUT 2005). The stakeholders in public education include parents, students, educational staff and administrators, government education officials, the academic community, and the general public, each with its own vision of the most vital indicators (Boland and Fowler 2000). Accommodating the needs and requirements of all stakeholders is almost impossible. If one selects too many indicators, data-collection costs will overburden the system, and a performance reward scheme will be hard to develop. If one selects too few, criticisms of incompleteness may arise, together with the possible consequence of "tunnel vision," which Smith (1995) described as the concentration of managers on a selected few quantifiable measures at the expense of the unquantifiable ones.

With the multitude of stakeholders of the educational system and a different definition of "educational performance" (or, for that matter, "quality") for each of them (Cullen et al. 2003, Jones 2003, Watty 2003, Tam 2001, Boland and Fowler 2000), the difficulty of producing quality education becomes clear. Educational achievement remains one of the most important indicators in evaluating educational systems (Alberta Learning 2002, Elmore and Rothman 2000, Fitz-Gibbon Kochan 2000, Haveman and Wolfe 1995, Windham 1988). However, a front-page article in the *National Post* newspaper pointed out the need to move "beyond standardizing testing as the way of measuring student and school performance" (Owens 2002, p. A1). In order to measure educational performance, educators now call for collecting "a mix of diverse and occasionally unlikely statistics" such as teen pregnancy rates, obesity and smoking rates, and volunteerism (Owens 2002, p. A5).

2.2.4. Organizational performance measurement frameworks in education

Educational organizations, like their business counterparts, benefit from using a consistent framework for measuring and reporting performance. Without consistency, too many or too few indicators can be used, input measures can be overemphasized, and concentrating on achieving academic excellence becomes harder to achieve. Organizing a set of indicators into a framework, such as a "Balanced Scorecard" (Kaplan and Norton 2001), allows the issues of academic, educational, and financial performance to be separated (O'Neil and Bensimon 1999).

Although assessing of the effect of performance measurement on the most important outcome of educational processes, namely the quality of teaching and learning, is difficult, the two benefits of developing a performance-measuring program in educational institutions are getting people to analyze their actions and their consequences, as well as encouraging communication among all the people involved in working toward achieving the established goals.

2.2.5. The Balanced Scorecard in post-secondary education

Cullen et al. (2003) advocated wider adoption of the Balanced Scorecard in higher education. The diverse interests of the many stakeholders in post-secondary education can be well represented by using the Balanced Scorecard's perspectives. Their interdependence will demonstrate that actions aimed at achieving one objective (e.g., increasing the student/instructor ratio to achieve fiscal targets) have consequences on the others (e.g., student satisfaction with the quality of education). The implementation of the Balanced Scorecard at a university may have an effect similar to the implementation of the Balanced Scorecard in a private organization – assuring of customers and partners about the quality of the organization's processes and outputs. The authors argued against the use of a generic Balanced Scorecard in a university department. Given the diversity of programs and organizations in higher education, the Balanced Scorecard has to be tailored to a particular department, faculty, or university. For example, the authors proposed a Balanced Scorecard for a faculty of management and business at a mid-ranking UK university. By examining the faculty's aims and objectives, both stated (e.g., academic research) and implicit (e.g., financial stability) Cullen et al. (2003) were able to create performance targets for each objective (e.g., publications, bursaries awarded).

O'Neil and Bensimon (1999) described the use of the Balanced Scorecard at the University of Southern California (USC) School of Education. Initially, the School perceived performance reporting requirements as a burden, and approached the performance measurement project from the position of "do it as quickly as possible and put the file away." Later, however, the faculty realized the value of performance information for receiving university resources and paid attention to the project. The faculty decided to organize performance data systematically and turned to the Balanced Scorecard, which the faculty perceived to be the appropriate framework for measuring performance of educational organization and processes.

The scorecard was adapted to the needs of an educational institution. The financial perspective was replaced with the academic management perspective, the customer perspective was renamed the "stakeholder perspective", and students and employers were identified as the major stakeholders, and the scorecard itself was named the "academic scorecard". As in public organizations, the "customers" (the students) were the most important priority, and the "quality of academic programs" was the most important indicator. In developing the measures, the School decided to keep the scorecard simple by having five or fewer indicators within each perspective. The benchmarks set for the indicators reflected the actions that were already under way at the School in order to improve its performance, which were the actions forced by the changes in environment, such as increased competition for the top students. The School felt that actions were producing the desired results but had no means to measure what exactly was working, and how well.

2.2.6. Measuring educational quality at the classroom level.

While a significant amount of effort and money is devoted to measuring educational quality, performance measurement in post-secondary education fails to reach the levels actually responsible for results – the levels of a department and classroom (Burke 2003). In addition to focusing on administrative quality measurement at the university level, continuous quality improvement must be ensured at the "educational-delivery level" (Jones 2003).

Among the many methods of evaluating educational quality at the classroom level, student ratings of instruction are one of the most widely used around the world (Harvey 2003, Griffin et al. 2003). Student evaluations of teaching effectiveness were introduced in several US universities in the 1920s. These evaluations' importance was fully recognized in the 1950s, when instructors realized that students will judge their effectiveness regardless of whether instructors want to acknowledge this fact, but that they can choose whether to use this information or not. Instructors also acknowledged that students are in the best position to report their own perception of their interest, attitude, and motivation, and, at the same time, to evaluate the quality of teaching (McKeachie 1990, Marsh 1987).

In his seminal work, Marsh (1987) has shown that students' ratings of instruction are a reliable, relatively uncontaminated, and useful source of information about teaching quality. These ratings are one of the best and most researched sources of information about educational effectiveness.

Students' rating-and-satisfaction data can be collected through a questionnaire in a "satisfaction survey" (Harvey 2003) form. Many such questionnaires have been developed. Among them are the Instructional Development and Effectiveness Assessment (IDEA), the Student Evaluation of Instructor (SEI), the Student Instructional Rating System (SIRS), the Students' Evaluation of Educational Quality (SEEQ), the Student Instructional Report (SIR), the Course Experience Questionnaire (CEQ) (Griffin et al. 2003, Jones 2003, Paswan and Young 2002), and probably many others, often of unknown origin (Marks 2000). The satisfaction surveys also often collect data on student performance, and these data are then analyzed to determine cause-and-effect relationships between factors such as satisfaction with the course and with the instructor, and academic performance (Griffin 2004, Kent and Hasbrouck 2003, Paswan and Young 2002, Marks 2000).

The collected data have several uses. Since universities are facing increased pressure to demonstrate educational quality in measurable ways (Jones 2003), student feedback on instructional quality facilitates accountability (Griffin et al. 2003, Harvey 2003). This feedback can be used to guide the internal continuous improvement process (Harvey 2003), and should be used by instructors to improve the quality of teaching (Marsh 1987).

Another, more controversial use of the student ratings is for the faculty's pay, tenure and promotion considerations (Griffin 2004, Harvey 2003, Marsh 1987). For several reasons, the faculty strongly resists such use:

- Few instructors in post-secondary education have formal teaching training (Marsh 1987);
- Lenient grading has been shown to positively associate with student ratings (Isely and Singh 2005, Griffin 2004), a fact that might undermine grading integrity. Additionally, instructors who merely entertain students may receive higher ratings even when these instructors' lectures have little substantial value (Wiers-Jenssen et al. 2002, McKeachie 1990, Marsh 1987);
- Even among academics, no agreed-upon criteria are available for measuring teaching effectiveness and student learning (Jones 2003, Marks 2000).

2.2.7. Additional use of student survey data

Universities can also use the student rating data for marketing purposes (Griffin et al. 2003, Harvey 2003). Some of the better-known college-rating guides utilizing student evaluations are *Canada's MacLean's' Universities Ranking Report* (*MacLean's* 2004), and *U.S. News & World Report America's Best Colleges* (*US News* 2005).

2.3. Monitoring educational processes

2.3.1. Pursuit of quality

The quality of education concerns everybody. Parents seek the best school for their first-grader, post-secondary students want the best possible education for their (or their parents') money, and the government wants to see the taxpayers' dollars spent on education wisely and efficiently. These demands places a responsibility on educators and educational administrators to demonstrate that their educational institutions – whether they are a pre-school educational facility, a school district, or a university – are capable of providing high-quality educational opportunities at a reasonable cost.

In order to reduce operating costs, increase productivity, or improve the quality of programs, many public and non-profit organizations now turn to the business sector in search of tools and solutions. Public post-secondary educational institutions have to compete with their private counterparts and with each other for government and students' money. In an attempt to gain a competitive advantage, universities are adopting "know how" developed initially for and by companies in the for-profit environment.

One of the major developments in quality assurance in the manufacturing industry was the adoption of ISO 9000 series of standards. Later, studies showed how these standards can be applied in the university environment (see, for example, Karapetrovic 1998, Karapetrovic et al. 1998). Universities have been compared to a production system, with incoming students as the raw material; student knowledge as the product created by the teaching and learning processes; faculty, facilities, institutional support and financial services as the resources; and students, employers, accreditation agencies, and society as the customers (Karapetrovic 2002).

Industry also possesses a variety of Statistical Quality Control tools used to analyze quality problems and improve production or service processes. These SQC tools have helped entire countries, most notably Japan, to achieve a competitive edge over their rivals. The development of SQC tools and methods is driven mainly by engineering faculties in universities. It was only fitting that the organizations that teach and research SQC started applying the same tools to control their own processes and to solve their own quality problems.

2.3.2. Statistical Quality Control tools in education

The same quality control tools used in industry can be applied in education. A classroom can be considered a mini production system, where a process of knowledge transfer from instructor to students takes place. If student knowledge is the variable of interest, factors outside of an instructor's control, such as natural ability, level of schooling, and time available for studying, will be responsible for the "natural" variability in students' knowledge. An instructor cannot easily eliminate causes of natural variability from the process of knowledge gain.

The assignable cause of variability in students' knowledge, from the in-class perspective, could be a poorly delivered lecture resulting in low knowledge gain for students. If this cause is not detected early enough, the students' low knowledge might manifest itself only on the final exam, and a larger than usual number of the students might fail. Powerful quality tools (such as the Ishikawa ("Fishbone") cause-and-effect diagram) can be used to discover the causes of poor performance after-the-fact, but a retrospective analysis may not be able to discover the true assignable cause if several confounding problems have led to an out-of-control situation. Even if the true cause can be found and a problem can be prevented in the future, much of the defective output (e.g., students' failure on a test) would have been produced already. If the problem is detected early, ideally as soon as it has occurred, it might be corrected by, for example, clarifying the subject in a subsequent lecture.

Pierre and Mathios (1995) used an attributes statistical control chart (also known as a *p*-chart) to record student performance in a summer mathematics and science program at San Jose State University. The data collection tool used in their study was called the "Monitored Assessment" (MA). Although twelve MAs were administered in total, only four were presented in the article. Each MA included a number of questions and was given to three students (n=3). Scores below 70 on a 0-100 scale were considered to be a "defect," and the proportion of "defective" scores was plotted on a *p*-chart.

Therefore, if all three students in a sample scored above 70, then $p = 0$, and if all three students scored below 70, then $p = 1$ (Pierre and Mathios 1995). The 12 samples of size 3 collected by authors were, probably, not sufficient to establish reliable control limits. Additionally, the small sample size produced the upper control limit above 1 and the lower control limit below 0 on the standardized *p*-chart, so no point could have plotted outside the control limits.

Besterfield-Sacre et al. (1998) described an application of non-parametric *p* and chi-square charts for monitoring enjoyment of math and science courses by first-year engineering students. For each data point, pre-survey responses were used to establish the control limits, and the post-survey responses were plotted on the chart. The data set had only four points, representing four academic years (93-94 through 97-98). To discuss the applicability of "run rules," the authors presented a hypothesized plot running through the 2000-2001 academic year. The applicability of such a plot is, however, questionable, for by the time a "run pattern" becomes obvious, two generations of students will have already graduated.

Karapetrovic and Rajamani (1998) described the application of the "Magnificent Seven" (Montgomery 1997a) quality control tools – a statistical control chart, flowchart, histogram, checklist, Pareto diagram, Ishikawa (or "cause-and-effect") diagram, and scatter diagram – in lectures, laboratory work, and tutorials at a major Canadian university. The authors illustrated how a systematic approach to quality control and improvement in educational organization can be applied to assure the stakeholders and customers of engineering education that a quality product – a qualified engineering graduate – is being produced. The proposed system of quality control tools can also be applied to satisfy accreditation requirements for engineering programs and faculties. The authors described how a tool called the "Modified Background Knowledge Probe" (MBKP), could be used to monitor students' knowledge gain during lectures.

Meijer (2002) studied the fit between a score pattern of an individual taking a computerized adaptive test, and the pattern predicted by item-response theory (IRT) models, in order to detect unusual response patterns. An "unusual" response pattern would be, for example, a test-taker answering difficult questions correctly and answering easy questions incorrectly, or answering questions randomly. To detect patterns of unusual behavior, the author used a CUSUM chart. The pattern predicted by an IRT model would be a mix of positive and negative residuals [x_i - p_i], where x_i is the binary (0,1) score on question *i*, and p_i is the predicted probability of correctly answering question *i*. A string of consecutive positive or negative residuals would indicate an unusual pattern. The author used the sum of residuals as statistics plotted on a CUSUM chart. An accumulation of positive or negative residuals would result in a CUSUM statistics plotting outside the control limits, indicating an unusual pattern of responses to the test questions.

Jenkins (2003) described the application of the quality control tools, most notably the statistical control chart, Pareto diagram, and histogram, at the Kindergarten – Grade 12 level of the educational system. In addition to using the quality control tools to monitor students' academic performance, the tools were also used to monitor such performance indicators as attendance, discipline infractions, and enthusiasm for learning. The collected data provided the information that was used to redesign curricula and instructional methods.

2.4. Discovering cause-and-effect relationships

When a problem is encountered, two general ways can be used to find a solution: direct manipulation of the actual system in which problem occurred, or studying a representation of the system, called a "model" (Kelton et al. 1998, Reklaitis et al. 1983).

Sometimes a system can be manipulated directly. By using statistically designed experiments, one can make purposeful changes to the system and observe and identify the reasons for changes in the system's output (Montgomery 1997b). For example, Abbot et al. (1990) carried out a factorial experiment to examine student satisfaction with various methods of collecting students' ratings of instruction. This approach has a definite advantage. If a parameter is changed and the output changes as well, the researchers know they are looking at the right thing (i.e., the real cause) (Kelton et al. 1998, Reklaitis et al. 1983).

However, in some cases, the "do first and see what happens" approach is not feasible. When an actual system does not exist, we cannot manipulate it. Experimenting with a system such as a chemical plant, or a nuclear power station might be prohibitively risky and expensive (the Chernobyl accident illustrated this point). In some cases, "playing with the system" is not ethically appropriate, such as in emergency response planning. Learning about the world by using real systems could be too time-consuming. Therefore, the second approach is to create a model of the system of interest (Kelton et al. 1998, Sterman 2000).

Appendix II presents an introduction to modeling theory. This introduction discusses mental models and their applicability, and describes the relationship between system complexity and modeling, mechanistic and empirical modeling approaches, data-collection methods, and model estimation.

2.4.1. Limitations of ordinary regression models

Regression models are widely used to analyze both mechanical and social systems. They also have their limitations because the implied simplifications of regression models frequently do not capture the important complexities of the real world.

The assumption that independent variables are mathematical variables measured with negligible error, and remain fixed at the levels pre-determined by a researcher, works well in physical sciences, where a researcher can have control over the independent variables. In the social sciences, experimentation with the independent variables is often infeasible, and the manipulation of their levels is often also unethical. If we wanted to establish a relationship between family income and students' academic achievement, would we intentionally place a sample of families into the "below-poverty" class, even if we could do so?

Therefore, in the social sciences, when we want to establish relationships among variables, observation is often the only data collection venue available (Jöreskog and Sörbom 1996a, Hox and Bechger 1998), but observational nature of data creates several complications that regression analysis is not well equipped to handle. Firstly, now all variables, independent and dependent, are subject to a random error and uncontrolled variation (Jöreskog and Sörbom 1996a).

An example will illustrate a second limitation. Imagine we want to measure the effect of time spent on homework on a student's test score. We want to test the hypothesis that the more time a student spends on homework, the higher this student's knowledge will be. How can we obtain data to test this assumption? We can ask students how much time they spend on doing their homework daily, and we can obtain the test knowledge via an oral or written exam.

But what contributes to the knowledge – the reported homework time or the real homework time? Lying on a question about homework time is not going to increase a student's knowledge, and there will be a reason – the true, not the reported, homework time influences the knowledge. However, the data we have are for the reported homework time. Do students always report the time spent on homework correctly? Even if a student has no reason to lie about his or her homework time, it is frequently reported in integer numbers of hours. Therefore, someone reporting to have spent 2 hours per day might have actually spent on average 1 hour and 55 minutes, or 2 hours and 15 minutes.

In addition, what does a test measure in reality? A test is intended to measure student knowledge, but how good is the test score as an indicator of student knowledge? Even if we had a perfect test, a student's performance on the test might not be perfect – illness, fatigue, or a personal situation can affect the student's performance. It is reasonable to assume that, at least to some degree, even a perfect test does not capture the true level of a student's knowledge.

All information we have about the real world comes via measurement. The data collected are only the indicators of the concepts of interest (Hayduk 1987). Only indicators (and not the concepts) can be observed directly, and all inferences about the unobserved real concepts must be drawn from the observed indicators (Hayduk 1987, Bartholomew 1983).

When dealing with a mechanical system, we frequently do not realize these facts because very often, the measurement error is negligible. When measuring the diameter of a shaft with a caliper, we can be relatively sure the diameter is 24.8 mm if the caliper's dial reads "24.8 mm." In social systems, however, many concepts cannot be observed directly (for example, student attitude), and we frequently rely on data provided by humans. Besides the inherent imperfection of measuring tools (e.g., interviews and survey forms), bias and imprecision in responses provide an additional measurement error.

2.4.2. From ordinary regression to Structural Equation Modeling

The summary statistics (e.g., the mean, variance, frequency diagram) can provide some useful information about an educational system and its variables, but in order to establish the nature of the relationship between system variables, either a designed experiment needs to be carried out, or observational data have to be collected and analyzed by using an empirical model.

Designed experiments, while mainstream in the technical world, are also sometimes carried out on human subjects (Simeonov et al. 2003, Pulat 1997), and in the field of education (Arias and Walker 2004, Abbot et al. 1990). Nonetheless, experimenting on human subjects is often impossible or unethical. If an improved lecture had to be tested, which students should be assigned to a control group listening to the old (and supposedly poorer) lecture? In such cases, an empirical model is estimated by using observational data.

Regression models are still widely used in the field of education (Isely and Singh 2005, Schultz et al. 2004, Griffin 2004, Wiers-Jenssen 2002, Stiefel et al. 2001). However, researchers start recognizing that models based on the analysis of variance and regression do not distinguish between unobserved concepts and indicators, as well as place causal effects improperly and do not account for indicator error variance (Paswan and Young 2002, Marks 2000).

Social scientists have recognized the deficiencies of regression analysis in dealing with the observational data and social systems and have developed a methodology known as "Structural Equation Modeling" (SEM). For a reader not familiar with this method of modeling, SEM basics (the origins of SEM, the development of SEM into a separate discipline, a SEM model in a pictorial and equation form, model estimation, and an assessment of adequacy of model fit) are presented in Appendix III.

2.4.3. SEM in Education

SEM is a versatile modeling approach. By using SEM techniques, one can carry out ordinary regression analysis, exploratory and confirmatory factor analysis, and specify and estimate structural causal models involving unobserved concepts and their observed indicators (Hox and Bechger 1998).

Factor analysis has been extensively used in the field of education. The first factor analysis studies of students' evaluation of teaching effectiveness were carried out in the 1940s. The quantity of research on student evaluations increased significantly in the 1970s, but the articles presented at conferences and published in journals were often of questionable quality (Marsh 1987). Researchers in the 1980s and 1990s tried to establish how different components of student surveys relate to each other and to other constructs, using factor analysis to demonstrate that a particular group of student ratings measure a separate and distinguishable characteristic of teaching effectiveness.

Marsh (1987) described several studies (including his own) in which factor analysis was used to select indicators of distinct components of teaching effectiveness. The factors identified in these studies fall into three general categories: instructor qualities (i.e., presentation clarity, interaction, enthusiasm), course qualities (i.e., organization, planning, workload, demands), and student characteristics.

Cranton and Smith (1990) used exploratory factor analysis to study how the structure of student ratings depends on the unit of analysis. The authors used individual students' ratings, class means, and the deviations of the individual ratings from the class means as units of analysis. Cranton and Smith (1990) illustrated that for the concepts "Interest / Atmosphere" and "Organization," the analysis based on the class means produced different groupings of indicators than analysis based on either individual ratings or deviations from the class means. The use of class means averaged out individual differences in how the students perceived instruction (interest and atmosphere), but perception of course organization remained consistent regardless of the unit of analysis.

Murray et al. (1990) studied how instructor personality characteristics contribute to positive or negative evaluations from students. Exploratory factor analysis was used to select indicators of instructor personality traits, and students' evaluations were used to assess instructor effectiveness in six types of psychology courses. The authors found that the perceived teaching effectiveness varied between different courses for the same instructor, that no uniformly effective or ineffective instructors had been identified, and that personality traits (i.e., liberalism, extraversion) contributing to effective teaching had different weights for different types of courses (i.e., undergraduate versus graduate).

While simple exploratory or confirmatory factor analysis is still used in education today (for example, Kent and Hasbrouck 2003), more advanced SEM models are being reported in the literature. For example, Marks (2000) created an SEM model based on student evaluations of instructor. This author discovered a number of causal effects, including a negative effect from the course workload/difficulty on the perceived fairness of grading, a positive effect from the perceived fairness of grading on the instructor's rating, a positive effect from the course organization on the instructor's rating and attitude toward the subject, positive effects from the attitude toward the instructor on the expected grade, and positive effects from attitude toward the course on the attitude toward the subject and the instructor's rating.

Paswan and Young (2002) studied the relationships among the factors "course organization," "student-instructor interaction," "course demands," "instructor involvement," and "student interest" in university courses in marketing. The authors postulated a number of hypotheses about the relationships among the variables, tested a model, and presented their findings: a positive reciprocal effect between instructor involvement and student interest, a positive effect of course organization on student interest, a negative effect of course demand on instructor involvement, and a positive effect of course organization on student interest. The authors also found no significant effect of course demand on student interest.

Grygoryev and Karapetrovic (2005) showed how an SEM model could be used to discover factors contributing to student knowledge (as measured by test scores) and attitude towards a course's subject. The model was based on the data from the Third International Math and Science Study (TIMSS) of 1995. The TIMSS study tested students in three age groups and from forty-two countries, in mathematics and science, and collected background data for students, instructors, and schools. The data used for the model consisted of student information relevant to science testing. The sample used in estimating the model's parameters consisted of 1,850 Grade 3 and 4 students from Alberta, Canada. The authors hypothesized that the concepts "Homework Time" and "Test Score" were involved in a reciprocal relationship. The model's estimates indicated that higher test results produced a more positive attitude toward the subject than lower test results. Higher test results, in turn, encourage a student to spend more time on homework. More time spent on homework produces even higher test results.

2.4.4. Limitations of SEM models

The SEM approach can be used as the first step for addressing a problem – gaining insight into a system. Can the SEM approach also be used to solve the problem? In the context of this research, SEM can be used to discover the causal structure of the relationships among the variables of a classroom educational system. Can the SEM approach be used to find ways to improve student performance and attitude toward a subject?

In the absence of other tools, using a SEM model is still better than modifying a system directly via a trial-and-error approach, but SEM models (and all regression models) have some limitations that make them less than perfect in designing and testing policies directed at solving the problem.

2.4.4.1. Assessment of fit versus assessment of causal claims

Model diagnostics focuses on the question "How well does the model fit the data?". Using a sample from population, we can estimate the model, but how do we validate it? In other words, how do we affirm that the model measures what it (or, rather, a modeler) claims to measure? How do we prove that the model represents a true causal world?

The question "Does correlation imply causation?" was the driving force behind the development of the path analysis method (Wright 1921), but ability of structural models to demonstrate causal relations still remains an issue of debate (Hayduk et al. 2003, Hox and Bechger 1998, Hayduk 1996). The debate is not SEM-specific, but is pertinent to the whole field of statistical modeling and estimation (Montgomery and Runger 1999, Levine et al. 1998).

The problem results from the use of observational, and not experimental data, in most SEM applications. According to Montgomery and Runger (1999), "designed experiments are the only way to determine causal relationships" (p. 446). When experiments on social systems are either impossible or unethical, researchers try to develope alternative methods of validating a model's causal claims. One approach is to use two data samples – one for model estimation, and one for model validation. The same sample can be split into two sub-samples (data splitting) and the methods in which the same sample is re-used include the bootstrap and the jackknife. Cooil et al. (1987) proposed a so-called simultaneous cross-validation method in which the same sample is used for model estimation subject to cross-validatory constraint.

Other methods consider introducing additional model constraints (see, for example, Hayduk et al. 2003 for a discussion of one such method, called *D-Separation*).

Some researchers consider that testing a model by using a second independent data sample acquired from either the same, or a different population, is the best model validation approach (Cooil et al. 1987).

2.4.4.2. Effect of time

Time, as a variable (or dimension), is not present in structural equation modeling, but is in the real causal world. Time delays between a cause and its effect are common in everyday life: how long does one poorly prepared lecture does take to undermine students' attitude toward an instructor? How much time is required for the instructor to regain the students' trust?

When issues of time are considered in statistical modeling, researchers typically assume that the effects introduced by variables occur instantaneously (Hayduk 1987). Since time is not present as a variable in SEM analysis, the value of an effect between two variables describes only a final change in the dependent variable, given a unit increase in the independent variable. However, this value does not tell us how fast the change happened, or by what process the system arrived at the dependent variable's final value.

Multiple regression or structural equation modeling implies that the output was obtained by maintaining inputs at a constant level for sufficient time for the system to settle in its output values (Heise 1975). One way to deal with time-varying inputs and time delays is to use cross-sectional research (Heise 1975). In this approach, the system output is measured on multiple instances by setting inputs at different levels. This approach also has limitations. Even when the inputs are controlled at steady levels, if the outputs are measured too early, a change in the dependent variables might not have occurred yet. Therefore, the researchers must wait until the effects take place and the system settles into its new values (Heise 1975).

2.4.4.3. Time-dependent inputs and feedback loops

When the inputs themselves vary with time, the system may never settle into stable new values. The inputs and outputs can still be observed at a given time, but if the exact nature of the system's structure (and its dynamic behavior) are not known, the outputs might not be able to be linked to the inputs that caused them (Heise 1975).

The presence of loops further complicates a situation. Loops create feedback, and feedback produces dynamic non-linear behavior in system variables (Forrester 1968). Depending on the signs of the coefficients connecting variables in a loop, a loop may be positive (also called "reinforcing" (Sterman 2000), or "amplifying" (Heise 1975)), or negative (also called "balancing" (Sterman 2000), or "control loop" (Heise 1975). A positive loop enhances the initial change in dependent variables included in a loop, while a negative loop counteracts the initial change.

For variables affected by loop or feedback effects, estimates of the structural coefficients connecting these variables cannot be considered as the overall effects of the variables on each other (Hayduk 1987). Hayduk (1987, 1996) provided matrix procedures for estimating the direct, indirect, and total effects for systems containing loops. While the ultimate numerical values of variables are of interest to a researcher, the processes by which a system arrives at those values is of equal or even greater importance. Variables involved in a complex relationship arrive at their final values non-linearly. A system might not be even able to achieve its final stage as its non-linear behavior might at times cause the system to become unstable and collapse (Heise 1975).

While several approaches have been suggested for dealing with time-varying inputs (such as averaging input observations, partitioning input signals into constant and variable components (Heise 1975)), a new discipline, called "System Dynamics," was developed for the purpose of studying the behavior of systems containing feedback (loop) structures over time.

2.5. Designing and testing policies by using System Dynamics

2.5.1. What is System Dynamics?

Regression models have limited use in analyzing systems with the technical characteristics of non-linearity, high-order complexity, and time dependence (Forrester 1968), and the social characteristics of "messiness, ambiguity, time pressure, politics, and interpersonal conflict" (Sterman 2000). System Dynamics (SD) is a method of studying socio-technical systems whose behaviour indicates the presence of feedback structures and the response of such systems to a policy change (Starr 1980, Jackson 2000). Jay Forrester developed SD in early 1960 at the Sloan School of Management of Massachusetts Institute of Technology. SD grew out of a one-page pen and paper simulation of inventory control system (Forrester 1989) into an applied methodology for dealing with complex non-linear socio-economic systems.

All systems can be divided in two classes: open and closed. Closed, or feedback, systems in turn can be divided into negative loop and positive loop systems (Forrester 1968):

- Open system: a system in which the output is determined by the inputs, but the output is isolated and does not influence the inputs. An open system cannot control its own performance and cannot adjust itself. An example is an automobile, which has no knowledge of where to go and is not controlled by where it has been.

- Closed system: a system in which an action is based on outcomes of previous actions. A closed loop brings the outcome of previous actions to a decision-making element to influence the future actions. A closed system can be either a positive-loop system, or a negative-loop system:
    - Positive loop system: a system that generates growth with action based on the results of previous actions. This system has no externally defined goal. An example is the growth of bacteria, for the new bacteria's growth rate depends on the bacteria accumulated from past growth;
    - Negative loop system: a system that has a goal and adjusts itself if it is not achieved. The controlled system parameter is subtracted from the goal and creates a discrepancy signal (Boland and Fowler 2000). An example is a house's heating system with a pre-set temperature and thermostat that adjusts the temperature when it falls outside the pre-set range.

Feedback loops are the main elements of a closed system. Two elements of a feedback loop are levels and rates. Levels accumulate flows in a system as the net difference between the inflow and outflow rates, and a rate describes how fast a level changes. The accumulation of flows in levels (stocks) creates the dynamic behavior of systems. Accumulation creates a "phase lag" between inflows and outflows, and the phase lag delays the instantaneous effect of the flow on other variables, creating a dynamic effect (Forrester 1968).

Negative feedback (which occurs when the current state of a system is compared to the desired state, and the difference is used to guide action) is of crucial importance, since the behavior directed towards the achievement of a goal depends on negative feedback. Communication is another important element in SD, since control involves the communication of information: the feedback loops in a system are connected by information links (Jackson, 2000).

SD involves constructing models of dynamic systems and testing them by using computer simulation. SD requires the combination of the human mind and computers. The human mind is important at the beginning of problem analysis – problem formulation, identification of feedback loops in the system, and at the conclusion – when the mind action to improve system behavior. The stage of simulating model behavior and testing different policies requires a computer since the human mind is weak when dealing with a complex problem exhibiting dynamic behavior (Jackson 2000).

SD, as a collection of tools, procedures, rules, and modeling methods, exists within the theoretical framework of Systems Thinking (SDS 2005). Appendix IV describes the System Thinking approach, its development into a cross-discipline, its structure, and the place of System Dynamics in it.

2.5.2. System Dynamics in education

Although System Dynamics has existed for about 50 years, it has not found wide application in the field of education, partially because educational systems, especially at the elementary school level, operate under legislative and judicial mandates, which are static in nature (Baker and Richards 2002).

Examples of educational SD models can be found in System Dynamics textbooks and reference texts. For example, Sterman (2000) described several dynamic models of the behaviour of an individual student. By using feedback structures, Sterman (2000) illustrated how a student's target level of achievement may fluctuate depending on his or her actual level of achievement. For example, if a goal is set too high and a student constantly underachieves, the student's frustration with his or her own abilities and with the academic system will increase. Frustration may lead the student to lower his or her expected level of achievement. Sterman (2000), however, developed cause-and-effect and stock-and-flow models without specifying model's equations and without simulating these models.

Richmond et al. (2000) described, modeled, and simulated the relationship between homework backlog and a student's workload. He illustrated that students who complete homework as it is assigned have no backlog and maintain a modest level of workload throughout a course. Students who spend little time on homework at the beginning of a course have a low level of workload initially, but will have a large homework backlog towards the end of the course. By trying to clear the backlog and working extra long hours, these students may "burn out."

Academics have investigated the applications of SD in education. Eftekhar (1998) used SD to study the processes of analysing and memorizing new information. By modeling the process of analysing and storing new information as a "main chain," he illustrated how information first is placed in the short-term memory, and then is either forgotten or is transferred into the long-term memory.

Salhieh and Singh (2003) analyzed how performance benchmarking and policy setting can be treated as a dynamic problem. The quality of teaching, research, and the students, at the level of a university department, were compared to the benchmarks, and the discrepancy between the perceived actual and the desired (benchmarked) quality served as the driver for policy change. Teaching quality was measured by two indicators: the percentage of students graduating within an "ideal" timeframe, and the percentage of students returning to university in the following semester. Research quality was measured by "gifts per faculty" and publications per faculty member. Student-body quality was measured by the mean entrance standardized test (such as SAT) score and the average grade point average. The authors illustrated, by using a SD model, how different admission policies (i.e., offers of financial assistance, various entrance requirements) and the allocation of funds between research and teaching would affect each of the performance indicators over time.

2.6. Summary

Chapter 2 presented a literature review on performance measurement, modeling, and management in education. Performance measurement originated in the business world, but found its way into post-secondary education as well. Surveys of student satisfaction with the process of instruction and with instructors is one way of collecting educational data in a classroom. The SEM approach can be employed to analyze the data in order to discover cause-and-effect relationships among the classroom educational system variables.

Educational processes occurring in a classroom, such as knowledge transfer between the instructor and the students, should be monitored on a continuous basis, since these processes are responsible for the educational outputs and outcomes. SPC methods used in the industry, such as the use of a statistical process control chart, may be useful in classrooms as well.

Since the classroom is a complex and dynamic environment, researchers must consider time and the feedback structure of a system when analyzing the effects of system structure and policy changes on the behaviour of key variables such as student performance and attitudes toward a subject.

Chapter 3: Motivation, Objectives, and Methodology

3.1. Introduction

This chapter begins by explaining the motivation for this research. Based on the problems identified in the literature review, the research objectives are established and the methodology for achieving them is described. The chapter also describes how a classroom is treated as a "system", and how a performance framework for measuring system inputs, processes, and outputs is developed.

3.2. Motivation

The motivation for this work came from the examination of the current state of the research on the subject of measuring, modeling, and managing engineering education at the classroom level. After reviewing the pertinent literature, it became obvious that the current research efforts are compartmentalized within particular academic fields and that solutions are sought for and implemented ad hoc. These problems create the need for a systematic integrated approach to resolving problems in a classroom in the post-secondary educational system.

3.2.1. Performance Measurement

Performance measurement is being increasingly wider accepted in the field of education. While the post-secondary institutions are often introducing performance measurement under competitive pressure, measuring educational outputs and outcomes is often a mandated activity in the Kindergarten – Grade 12 education system (Alberta Finance 2001, OPB 2001).

Some authors had remarked that, to the detriment of the system as a whole, educational professionals concentrate their attention and efforts only on those areas that are measured (Taylor 2001, Smith 1995). Therefore, to ensure that all aspects of an educational system were being covered, educational organizations turned to the for-profit sector in search of performance measurement frameworks. One of the best-known frameworks, the Balanced Scorecard developed by Kaplan and Norton in 1992, has been applied in higher education (Cullen et al. 2003, Chang and Chow 1999, O'Neil and Bensimon 1999). This application has been limited, as only a few universities are using the Balanced Scorecard.

Another deficiency of the current performance measurement practices is their failure to reach the level actually responsible for producing results – the classroom itself (Burke 2003, Jones 2003).

Currently, at the classroom level, data are collected in the form of student evaluations of the instructors' teaching effectiveness. While many tools are available, researchers in higher education, nonetheless, have long designed their own instruments to collect data on student performance, satisfaction, and attitudes. (Isely and Singh 2005, Griffin 2004, Harris and Bretag 2003, Kent and Hasbrouck 2003, Worthington 2002, Marks 2000, Abbot et al. 1990, Marsh 1987). Researchers have had several reasons for doing so:

- Agreed-upon criteria for measuring teaching effectiveness and student learning are lacking (Jones 2003, Marks 2000, Marsh 1987);
- Student satisfaction surveys are sometimes used primarily for administrative control rather than for ensuring continuous educational quality (Harvey 2003);

- When institutional tools are used, instructors may have access to the data summary only and cannot modify the tools for their own research-specific needs (Jones 2003).

To construct, for example, an SEM model, a researcher needs at least a matrix of covariances among the performance indicators. Normally, the only data available are a descriptive summary statistics (e.g., mean, standard deviation) for each indicator, and the placement of the mean relative to that of similar courses. Additionally, department-wide surveys are usually conducted at the end of a semester, so any course improvements can be implemented only for the next semester.

A more serious deficiency of student-satisfaction surveys and student ratings of instruction is that neither instrument directly measures the process of learning (Wiers-Jenssen *et al.* 2002). Only the outcome – the student's final course grade or, more often, the expected final course grade – is available from either of the tools. Teaching effectiveness is also not measured directly and quantitatively, but indirectly (as the students' perception of effectiveness) and qualitatively (comparatively to the effectiveness of other instructors).

3.2.2. Process Control

Despite their usefulness and rich history of industrial application, SQC tools are not widely used in educational institutions. When they are, these tools are not institutionalized organization-wide, but used at the instructors' own initiative. Only a few reports have been published on application of SQC tools in the field of education (Grygoryev and Karapetrovic 2005, Meijer 2002, Karapetrovic and Rajamani 1998, Besterfield-Sacre et al. 1998, Pierre and Mathios 1995).

While industry has largely moved away from inspecting the final product to trying to control the prosesses (Montgomery 1997a), educational organizations still concentrate primarily on controlling outputs (such as, for example, graduation rates) and outcomes (such as, for example, test results), while ignoring, or paying little attention, to the processes responsible for producing outcomes and outputs (Tam 2001). While an intervention to improve quality is most effective at the level where the processes actually take place (i.e., a shop floor in a manufacturing plant) (Montgomery 1997a), at the level of classroom – the level where education actually occurs – process control is almost non-existent in educational systems.

Because of the sparse literature on the topic, research on the use of quality engineering tools in educational processes is much needed. Dr. Norman Fortenberry (1999), Director of the Undergraduate Education Division, National Science Foundation (NSF), identified the application of quality control to influence instructional outcomes as one of the methods that provides the basis for high leverage priorities in engineering education research.

3.2.3. System Modeling

The educational SEM models reported in the recent literature have two serious methodological problems:

- Their exploratory approach to model construction
- Their use of "close fit" indexes to justify ill-fitting models.

Factor-analytic models have been employed in education for a long time (Kent and Hasbrouck 2003, Cranton and Smith 1990, Murray et al. 1990, Marsh 1987,). Factor analysis is used to select a group of indicators that best measure some underlying concept (e.g., "use of lecture time," "clear explanation of purpose of lecture," "use of multi-media tools" as indicators of the concept "lecture organization").

Factor analysis is frequently conducted in an exploratory way – a researcher rotates combinations of indicators corresponding to particular concepts, until model fit is achieved. This approach amounts to what Levine et al. (1998) called "data snooping" – examining data patterns before creating one's own hypothesis. Data, therefore, are used to guide the researcher's theory instead of to test it. Hayduk (1987) stated that the data used to adjust a model should not also be used to test it. Adjusting a model to fit the data could lead to capitalizing on the sample's random properties (Hox and Bechger 1998).

Confirmatory factor analysis is used *a priori* to test models instead of selecting indicators. Still, the deficiency of either exploratory or confirmatory factor analysis is in the lack of specification of a particular causal relationship structure among the concepts measured by the indicators. Researchers simply allow the concepts to freely co-vary with each other.

When the causal structure among the unobserved concepts is specified, researchers sometimes carry out the factor analysis to select the "best" indicators for the concepts (Moore 2005, Hahs-Vaughn 2004, Wittmann and Hattrup 2004, Paswan and Young 2002, Marks 2000). Hayduk (1996) and Hayduk and Glaser (2000) provide a comprehensive critique of this so-called "two-step" approach.

The second problem, arguably more severe, is researchers' inability to produce fitting models. When a researcher presents conclusions based on a model's results, the model must represent the real world adequately. The adequacy of SEM models is best assessed by using the chi-square test, which compares a model-produced matrix of covariances among observed indicators to a similar matrix obtained from the real world (Hayduk 1987). Statistically significant chi-square values indicate that the difference between the model-implied and the observed data sets amounts to more than the random fluctuation, and that some real-world causal forces have not been accounted for. The rule-of-thumb level of significance of 0.05 for the test statistics is normally used as a rejection/non-rejection criterion (Levine et al. 1998), while levels as high as 0.20 have been advocated for use in SEM (Hayduk 1987). Meanwhile, some authors reported that their models achieved "good fit" with *P* values as low as 0.013 (Kent and Hasbrouck 2003, Pang 1996), 10^{-97} (Geor[illegible] and Kaplan 1996), and even 10^{-274} (Paswan and Young 2002). Some authors (Hahs-Vaughn 2004) even did not report the degrees of freedom and the *P* value at all, leaving the reader to try to figure out the true level of the statistical significance of the chi-square test statistic (the best-case scenario would result in $P = 10^{-45}$ – was there a reason for not reporting it?)

Despite the fact that the above-mentioned models failed the chi-square test of an exact fit, the modelers still claimed that their models provided a good fit to the real-world data. These modelers were able to do so by using so-called indices of "close fit," such as the chi-square per degree of freedom, the root mean square error of approximation, and some others (for a detailed description of the approaches for assessing "closeness of fit", see Browne and Cudeck 1992). Effectively, these modelers were testing whether their ill-fitting models were "close enough" to the covariance matrix and whether small, but significant model misspecifications could be disregarded. Indeed, even some statistical textbooks call for looking at the practical significance in addition to the statistical significance of test results (Montgomery and Runger 1999). Nonetheless, acceptance of ill-fitting models as adequate on the basis of the "practical insignificance" argument may lead to the failure to discover causal factors of great importance. Hayduk et al. (2005) demonstrated how disregarding seemingly "negligible from a practical point of view" (but statistically significant) signs of ill-fit may have obscured the presence of a factor affecting ability of human natural killer cells to fight cancer.

3.2.4. System Analysis

System Dynamics methodology was developed for studying the behavior of systems containing feedback (loop) structures, over time and was prompted by the inability of applied methods such as Operations Research, Mathematical Optimization, and Linear Programming, to resolve problems arising in systems that are non-linear, time-dependent, and have high-order complexity.

Often, though, SD models are created on uncertain foundations. Some modelers adjust model parameters until historical behaviour is replicated (Pavlov and Saeed 2004). Some insert model parameters without justifying where the parameters came from and why the particular values were selected (Arquitt and Johnstone 2004). Some proceed to "what-if" and hypothesis analysis without validating their models with the real-world data (Dudley 2004, Baker and Richards 2002).

Some modelers stop at formulating a dynamic hypothesis without actually testing it (Cavana and Mares 2004). While a mere description of a system may improve understanding of a problem, system behaviour cannot be deducted reliably by just looking at the causal diagram. To solve a real problem, the real quantitative data are required (Coyle 2000), but even when they used, sometimes the way in which the model parameters have been estimated is not reported (for example, Homer et al. 2004).

In order to observe model behaviour that has real-world implications, we need real-world data. The quantification of models through "guesstimate" may actually produce less reliable results than pure qualitative analysis (Coyle 2000). The better techniques are used to estimate the relationships between a model's elements, the more confidence one can have in the model's results.

3.3. Objectives

The following list presents a summary of objectives of this research:

- Analyze an engineering classroom from the systems perspective in order to define classroom inputs, processes, and outputs / outcomes;
- Identify a set of performance indicators for the selected inputs, processes, and outputs / outcomes;
- Design a tool for collecting data on selected inputs and outputs / outcomes;
- Design a tool for collecting data on selected classroom processes;
- Develop procedures for continuous monitoring of the educational processes responsible for student learning at the classroom level;
- Postulate and test a cause-and-effect model to discover the factors contributing to such educational outcomes as student performance and attitude toward the subject, by considering
    - Observational character of data collection;
    - Measurement error and bias.
- Identify opportunities for improving the delivery and content of the undergraduate engineering management course;
- Design, test, and implement policies aimed at improving educational processes, output, and outcomes in an engineering classroom, by taking into consideration
    - The feedback structures present in the classroom educational system;
    - The effect of time.
- Evaluate effects of the implemented policies, and identify opportunities for continual improvement.

3.4. Methodology

The methodological approach employed to achieve the research objectives, and the chapters of the thesis in which a particular issue is addressed, are presented in Figure 3.1.

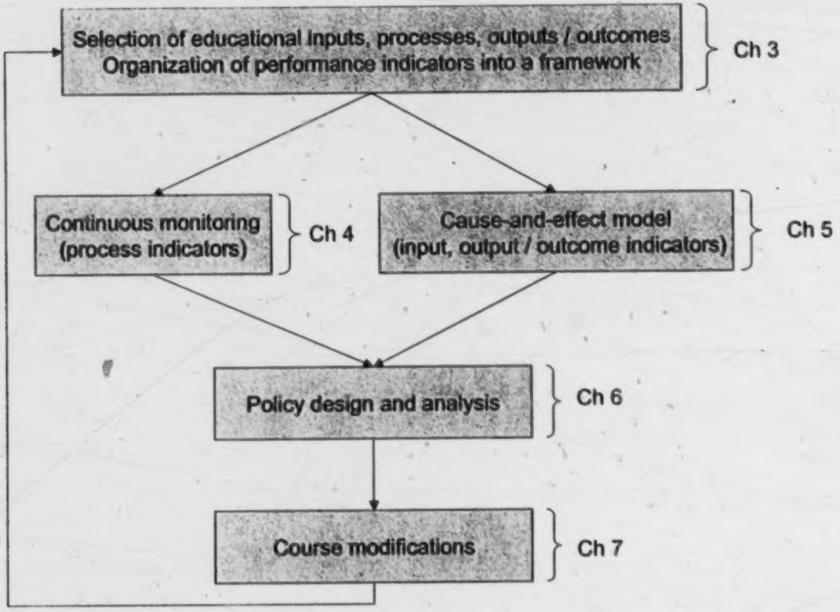

Selection of educational inputs, processes, outputs / outcomes
Organization of performance indicators into a framework
Ch 3
Continuous monitoring (process indicators)
Ch 4
Cause-and-effect model (input, output / outcome indicators)
Ch 5
Policy design and analysis
Ch 6
Course modifications
Ch 7

Figure 3.1: Research methodology

This work started with several questions in mind:

- What can / needs to be done to improve instructor's teaching effectiveness and students' learning effectiveness in a classroom?
- How do we establish what exactly affects students' performance, attitudes, and satisfaction in an engineering classroom?

- What is a safe way to test the effect of changes in the course content and delivery on performance indicators?

Figure 3.1 outlines how these questions will be answered.

A classroom educational system will be analyzed to determine what inputs and processes are responsible for educational outcomes and outputs. The performance indicators for the inputs, processes, outputs, and outcomes will be selected and organized into a performance framework. The methods of data collection will be established for each kind of indicators, and the performance data will be collected.

It was argued in previous chapters that controlling processes responsible for the output is more effective than inspecting it. An approach based on Statistical Process Control will be employed to control educational processes and to detect situations when the process goes out of statistical control. A different approach, based on Structural Equation Modeling, will be used to discover the cause-and-effect relationship between the input and output variables in a classroom educational system. The dynamic behaviour of the model, produced by feedback structures, will be analyzed by using the System Dynamics approach.

The insight gained from the process control and cause-and-effect model will be combined with the qualitative feedback provided by the students to design improvements to an undergraduate engineering management course. The policies believed to improve the system's behaviour will be implemented, and their effects will be evaluated.

The steps described above will serve as a basis for the continuous improvement effort. The implementation of changes to a course may require selecting new performance indicators. The variables initially left outside the model's boundary may need to be included in the feedback structures in order to enhance understanding of the classroom dynamics. After the selected processes are brought under control and are monitored on a continuous basis, new candidates for analysis may emerge. Once the effects of the introduced changes are assessed and understood, the cycle may begin again.

3.5. Performance framework

3.5.1. Classroom as a system

One of the fundamental quality improvement principles is that the desired results are achieved most efficiently when activities and related resources are managed as a system (ISO 9000 (2000)). Jenkins (2003) advocated application of Dr. W. Edwards Deming's quality improvement philosophy and argued in favour of the systems approach to managing education. The author defined a "system" as "a network of components within an organization that work together for the aim of the organization" (Jenkins 2003, p. 23). The goal of the K – 12 educational system is then "to produce quality high school graduates" (Jenkins 2003, p. 22). The educational system, in Jenkins' interpretation, consists of seven components: aim, customers, suppliers, input, process, output, and quality measurement (Jenkins 2003).

A "system" can also be defined as a set of processes that utilize resources to achieve an objective (Karapetrovic 1998). Karapetrovic (1998) and Karapetrovic et al. (1998) described how a university could be treated as a system. A similar, but simplified, approach can be employed to consider a classroom as a system. A "classroom system," then, can be defined as a set of educational processes occurring in a classroom that use educational inputs to achieve desired educational outputs and outcomes.

One of the approaches for defining classroom educational inputs, processes, and outputs is to analyze what students and instructors bring to the classroom (i.e., inputs), what students and instructors do in the classroom (i.e., processes), and what students and instructors accomplish (i.e., outputs and outcomes) (Marsh 1987).

Industry employs different methods of controlling inputs and outputs than the methods used to control processes. For example, in manufacturing, acceptance sampling is used for input control (and, since the output of one process is the input for another, acceptance sampling can be considered as the input/output control method), and various SPC methods are used for controlling the processes. Statistical modeling, in the form of ordinary regression analysis, or in the form of SEM, can be used to establish the causal relationships between the inputs and the outputs, while treating processes as a "black box."

The systems approach to performance measurement in a classroom can be broken down into the following steps:

1. Defining system boundary – determining which classroom inputs, processes, and outcomes / outputs will be included in analysis;
2. Separating system elements into inputs and outcomes / outputs, and processes – since different methods are used for measuring and controlling each element;
3. Determining the data requirements – what should be measured, how it should be measured, and how often;
4. Designing / selecting data collection tools;
5. Designing / selecting data analysis tools;

Figure 3.2 illustrates the classroom performance framework based on the systems approach.

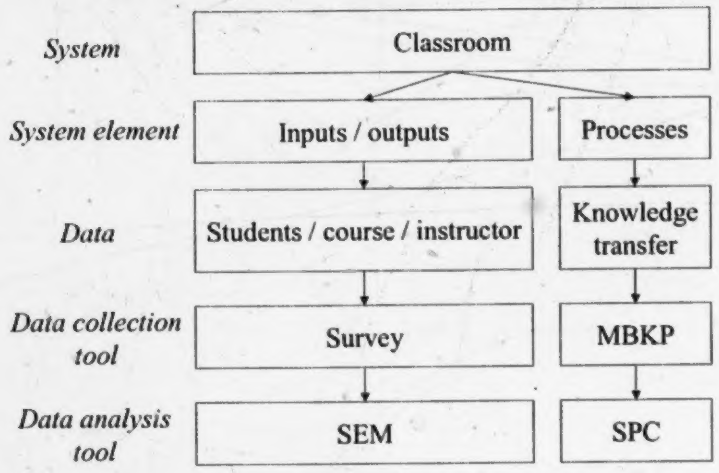

System
Classroom
System element
Inputs / outputs
Processes
Data
Students / course / instructor
Knowledge transfer
Data collection tool
Survey
MBKP
Data analysis tool
SEM
SPC

Figure 3.2: Classroom performance framework based on the systems approach

3.5.2. Inputs, outcomes, and their indicators

The inputs and outputs / outcomes for a classroom system can be arranged in a framework in different ways. One way is to group them along the "student-instructor-course" lines. Eftekhar (1998) used a teaching/learning system consisting of three major components: student characteristics, instructor characteristics, and subject matter characteristics. The input / outcome characteristics of a university system presented by Karapetrovic (1998) can also be divided into the students – instructor – course elements.

Student surveys are often conducted to collect students' opinions on the quality of instruction and to allow students express their attitudes toward the subject, the instructor, and the course. The literature review indicated that, despite some controversy, the use of student surveys and ratings of instructional quality is a reliable and valid method of collecting information on student attitudes and satisfaction. Early research indicated that the effect of student background / demographic characteristics on student evaluation of instructors was minimal (March 1987). Recent literature indicated that such factors as student background (e.g., age, gender, ethnic origin), as well as the instructor's physical characteristics (gender, minority status) might influence student ratings of teaching effectiveness (Worthington 2002). Therefore, surveys today often contain questions about the students' background characteristics, to statistically control for their effect on evaluation.

It was decided to design a survey to collect indicators of educational inputs and outputs, and students' background characteristics. The survey was modeled on five student evaluations of instruction tools presented by Marsh (1987). Analysis of the tools indicated that the survey questions were grouped into the three general categories of student, instructor, and course characteristics.

3.5.3. The survey

Table 3.1 presents the elements of the survey designed to collect data on student inputs, outcomes, and background characteristics. This survey was given to the students taking an undergraduate engineering management course at the University of Alberta (the details of the survey and its administration are presented in Chapter 5). The data on the course and instructor characteristics were available to the author directly, since the author lectured in the course and also had participated in designing some of its elements (assignments, midterm exams, and some lectures).

3.5.4. Educational processes

While measuring educational outcomes, such as attitude toward the subject or performance on a test, is important, the educational processes actually produce the outcomes. Since many processes may take place simultaneously in a classroom, it was decided to limit the scope of this research, in order to concentrate on, arguably, the most important process responsible for student learning in an engineering education – the process of the transfer of knowledge from the instructor to the students in a classroom.

Table 3.1: Student inputs, outcomes, and background characteristics

Parameter	Input / outcome, background	Performance indicator
Financial background	Background	Owning a personal computer
Gender	Background	Reported gender
Language background	Background	Reported language background
Language practice	Background	Speaking English in everyday life
Age	Background	Reported age
Enrollment level	Background	Enrollment level in the course
Importance of having fun while in university	Input	Reported importance of having fun in university
Importance of succeeding academically	Input	Reported importance of doing well in university
Academic background in the discipline	Input	Taking a similar course previously
Time devoted to self-studying	Outcome	Reported homework time
		Reported time spent reading text/notes
Attendance	Outcome	Reported number of lectures missed
Student knowledge	Outcome	Midterm test score
Perceived course workload	Outcome	Reported perceived workload
Satisfaction with the course in general	Outcome	Reported satisfaction with the course
Satisfaction with the instructor	Outcome	Reported satisfaction with the instructor
Attitude toward the course subject	Outcome	Reported attitude toward subject
Balancing academic performance and recreation while at university	Outcome	Report on possibility of doing well and having fun while at university
Extra-curricular activities	Outcome	Participation in sports and cultural events

3.5.5. Modified Background Knowledge Probe – a tool to measure knowledge transfer

Angelo and Cross (1993) described fifty Classroom Assessment Techniques (CATs) designed to measure student knowledge. Some of these CATs are relatively simple and can be used during every lecture, while others, more complex, can be applied only once per semester (Angelo and Cross 1993). Karapetrovic and Rajamani (1998) and Karapetrovic (2002) illustrated how one of these CATs, the Background Knowledge Probe (BKP), can be modified to collect data on student learning in a classroom setting. The new tool was called a "Modified Background Knowledge Probe" (MBKP).

Using a BKP, the instructor asks one or more questions (reflecting important issues to be covered during a lecture) before the lecture (Angelo and Cross 1993). Using an MBKP, students answer such questions once prior to the lecture, and then again after the lecture. An MBKP may contain several "Before and After" questions. Each question corresponds to a particular important concept covered during the lecture for which the MBKP is prepared. If, for example, two important concepts are introduced during a lecture, then the MBKP for the lecture would have two "Before and After" (B&A) questions. Each "B&A" question has a number of answers (usually 3 to 5) in a multiple-choice format.

A typical MBKP administered during an undergraduate engineering management course at the University of Alberta is presented in Figure 3.3. MBKPs, each containing a number of B&A questions, are distributed to the students at the beginning of a lecture. The students answer these questions for the first time before the lecture and for the second time after the lecture by writing the letter corresponding to the selected answer in the box marked "BEFORE" and "AFTER" below the question, respectively. The details of MBKP administration, data collection, and analysis are provided in Chapter 4.

The MBKP was designed not as a grading tool, but as a student-learning measurement and improvement instrument. The instructor using an MBKP should emphasize that the data collected will not be used for grading purposes, but for course improvement only. Doing so should discourage students from guessing and, therefore, from negatively affecting the true value of the statistics collected. To reduce guessing, every question should include the "I do not know" option (Karapetrovic 2002).

When any sort of a test is used to assess the quality of instruction and not as a grading tool, students might not take the process seriously and might answer questions by guessing, thus introducing bias into the answer pattern (Meijer 2002). To maintain the students' interest in the process of measuring classroom knowledge transfer, the instructor can give the students frequent feedback by providing the answers to the B&A questions, discussing the results, and presenting up-to-date control charts with the plotted statistics.

ENGG 401 BEFORE AND AFTER LECTURE QUESTIONS (Jan. 16, 200X)

13. Which of the following entries from the income statement is included on the statement of retained earnings?

a. Operating Income
b. Net Income
c. Earnings Before Interest and Taxes (EBIT)
d. I don't know

ANSWER:	BEFORE	AFTER
	☐	☐

14. If in year 2001, your Net Income was $100,000 and you paid $120,000 in dividends, the Retained Earning as of December 31, 2001 would be

a. Greater than Retained Earning as of December 31, 2000
b. Same as Retained Earning as of December 31, 2000
c. Smaller than Retained Earnings as of December 31, 2000
d. I don't know

ANSWER:	BEFORE	AFTER
	☐	☐

15. A small business will be most concerned with achieving

a. Book break even
b. Cash break even
c. I don't know

ANSWER:	BEFORE	AFTER
	☐	☐

16. You realized that depreciation in your company was significantly underestimated for the upcoming year. In the light of this new information, it will be harder for your company to achieve

a. Book break even
b. Cash break even
c. Both book and cash break even
d. Neither book nor cash break even
e. I don't know

ANSWER:	BEFORE	AFTER
	☐	☐

Figure 3.3: MBKP (Example)

3.6. Limitations

A college classroom is a complex environment where many processes take place simultaneously, converting various inputs into outputs and outcomes. The scope of this research was limited to a number of classroom educational inputs and outputs, and one educational process occurring in a classroom. The goal of this research was not to build a template that can be applied in any environment. The specific variables, indicators, and constructs produced by this research may not be transferable to every classroom. Critical analysis is necessary in applying the proposed models under different circumstances, but the systems approach to measuring, modeling, and managing the classroom processes is common enough to be applied by anyone interested in improving classroom performance.

3.7. Summary

This chapter presented the research motivation, the research objectives, and the methodology to be used in accomplishing the objectives. A classroom was treated as a "system," and the systems approach was used to identify classroom inputs, processes, and outcomes. A performance framework based on the input / outcome indicators and the process indicators was constructed. A student survey was proposed as a tool to collect student input and outcome data and background characteristics. A Modified Background Knowledge Probe was introduced as a tool to measure the knowledge transfer between the instructor and the students in a classroom. Chapter 4 provides details on data collection using an MBKP, and data analysis using the SPC tools.

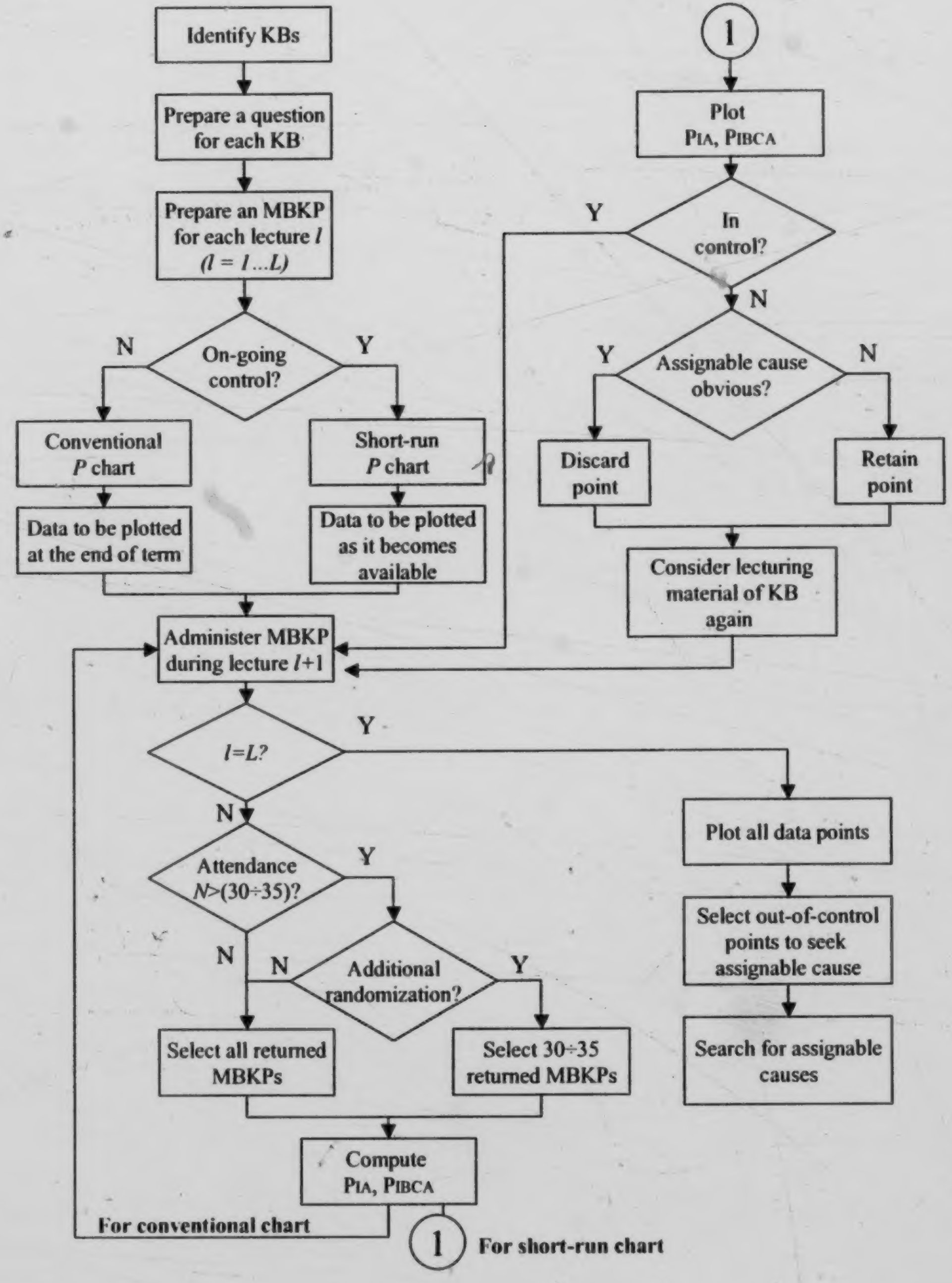
Chapter 4: Control of Educational Processes

4.1. Introduction

This chapter presents a tool which an instructor can use to measure on an on-going basis the students' knowledge gain as it occurs in a classroom, i.e., during the student-instructor interaction. The model described here can be used independently by an individual instructor to evaluate his or her contribution to the student learning in a classroom (which indirectly reveals the quality of teaching as well), or this model can be implemented department-wide to help in meeting a school's accreditation criteria, such as those of the Accreditation Board for Engineering and Technology (ABET), or the Canadian Engineering Accreditation Board (CEAB).

4.2. Implementation

An algorithm describing the design and administration of the MBKP tool is presented in Figure 4.1. The instructor should prepare one MBKP for each lecture during which student learning is to be evaluated. A whole set of MBKPs can be created prior to the beginning of a semester, or individual MBKPs can be prepared prior to each lecture.


Identify KBs
Prepare a question for each KB
Prepare an MBKP for each lecture l (l = 1...L)
On-going control?
N
Y
Conventional P chart
Short-run P chart
Data to be plotted at the end of term
Data to be plotted as it becomes available
Administer MBKP during lecture l+1
l=L?
Y
N
Attendance N>(30÷35)?
Y
N
Additional randomization?
N
Y
Select all returned MBKPs
Select 30÷35 returned MBKPs
Compute PIA, PIBCA
For conventional chart
1
For short-run chart
1
Plot PIA, PIBCA
In control?
Y
N
Assignable cause obvious?
Y
N
Discard point
Retain point
Consider lecturing material of KB again
Plot all data points
Select out-of-control points to seek assignable cause
Search for assignable causes

*KB – Knowledge Block

Figure 4.1: Algorithm for the design and administration of MBKPs

4.3. Statistics computed from an MBKP

Each B&A question has four possible answer combinations (see Table 4.1):

Table 4.1: B&A answer combinations

Before	**After**	**Statistic Collected**
Incorrect	Incorrect	Proportion of "Incorrect After", P_{IA}
Correct	Incorrect	
Incorrect	Correct	Proportion of "Incorrect Before, Correct After", P_{IBCA}
Correct	Correct	Proportion of "Correct Before, Correct After", P_{CBCA}

The rationale behind using each statistic is as follows:

- The proportion of "Incorrect After" includes the students who left the class without being able to answer the question and, presumably, did not benefit from the lecture and/or did not learn its main points. Here, the word "presumably" is used intentionally, since another possible explanation is that the students did not understand, or were confused by the question itself (an algorithm for testing whether poor lecture quality or poor question quality caused the problem is presented in section 4.8). The proportion of "Incorrect After" statistic includes both the "incorrect before, correct after" and "correct before, incorrect after" answers. If the latter number is consistently small, it can be considered an anomaly, and we can assume that the lecture was not the cause of confusion. If this number is consistently high, the instructor needs to question both the quality of his/her lectures and the quality of the B&A questions. In the data presented herein, the proportion "correct before, incorrect after" was very small, and we did not track this number as a separate statistic.

- The proportion of "Incorrect Before, Correct After" includes the students who did not know the answer to the question before the lecture, but did know after. This statistic is supposed to measure the learning that occurred in the classroom as a result of the lecture delivered by the instructor (and, thus, reflects on the instructor's teaching effectiveness).
- The proportion of "Correct Before, Correct After" includes the students who knew the answer both before and after the lecture. A high P_{CBCA} is not necessarily a desirable situation, since it indicates that many students did not learn anything new and did not benefit from the lecture.

Only two of the three statistics above need to be collected, since the third one can always be derived because the sum of all three statistics equals one:

$$P_{IA} + P_{IBCA} + P_{CBCA} = 1. \tag{4.1}$$

4.4. Use of the collected data

With the data collected from MBKPs, an instructor will be able to answer questions such as "How effective does the transfer of knowledge during the lecture seem to be?" and "What proportion of students seem not to have benefited from the lecture?"

P_{IBCA} provides a measure of the "knowledge transfer" process. Theoretically, the most desirable value for this measure is 100%, which would be obtained if none of the students knew the material prior to the lecture and all knew it after. However, several factors may prevent the P_{IBCA} from reaching 100% in practice. First, some students, even among the undergraduates, are likely to have previously taken courses in similar areas. In the example of the introductory engineering management course described in this chapter, a portion of students had prior exposure to the basic principles of management, finance and economics. For this group of students, no knowledge gain might occur during at least some of the lectures (their answers to the B&A questions may fall into the "correct before, correct after" category). Second, in most classes, and especially in larger ones, a mix of different learning styles is inevitably present. Those students whose learning style does not fit the instructor's teaching style may not fully benefit from the lectures and may require additional self- or assisted study. In this case, the "incorrect after" answers to B&A questions are expected. The third challenge to the assumption of the stability of the knowledge transfer process is that some topics are harder to learn than others. For instance, in engineering management, students frequently struggle with the concepts of break-even and depreciation. Finally, poor teaching, not learning, may be causing the problem.

P_{IA} provides a measure of the number of students who failed to learn material during a lecture. The theoretical goal for this value is 0%, as one would ultimately want every student in every lecture to know the correct answer to every question. However, as quality control practitioners know, the zero percent "defective" output rarely occurs in practice.

4.5. Example

4.5.1. Data source

The data presented here come from the administration of MBKPs during four semesters of an undergraduate engineering management course taught by three different instructors at the University of Alberta. Student participation in the study was voluntary, and anonymous. Since the research involved human subjects, an application was made to, and was approved by, the University of Alberta Research Ethics Committee (see Appendix V). Six sets of MBKP data were collected in total (see Table 4.2):

Table 4.2: Description of MBKP data sets

Set	Term	Instructor
1	Winter 2002	A
2	Fall 2002	A
3	Fall 2002	B
4	Winter 2003	A
5	Winter 2004	A
6	Winter 2004	C

In data sets 1, 2, and 3, the course content and lecture notes were similar, and the same MBKPs, designed by Instructor A, were used. In set 4, the MBKPs were designed by instructor C, while the course content stayed similar to that covered by the sets 1 and 2. In sets 5 and 6, different sets of MBKPs were used, and the course content became quite different between the two sections.

For analysis purposes, only sets 1, 2, 3, and 4 will be used here. Sets 5 and 6 are not comparable between themselves due to the large number of different factors (e.g., course content, MBKP questions) contributing to the effect of the knowledge transfer process in addition to the primary parameter of interest – the instructor.

The B&A questions used in MBKP sets 1, 2, and 3, are presented in Appendix VI, and the summary of the data collected from MBKPs is presented in Appendix VII. Set number 4 had 14 (out of 28) B&A questions different from those in sets 1, 2, and

3, and those different questions are marked with an asterisk in Appendix VII. For convenience, the summary of the data is presented below in Table 4.3.

Table 4.3: MBKP data sets used in the analysis

Data set	Instructor	Number of questions	Total # of observations	Estimated process P_{IBCA}	Estimated process P_{IA}
Winter 2002	A	28	1612	0.276	0.189
Fall 2002	A	28	1484	0.262	0.195
Fall 2002	B	28	2636	0.180	0.311
Winter 2003	A	28	1836	0.202	0.215

4.5.2. Statistical principles

The quality characteristic chosen for monitoring knowledge transfer during a lecture – the students' answer to B&A questions – was measured on the "correct/incorrect" basis, rather than assigned a numeric score, say, out of 100. All instructors are probably familiar with such constructs – they routinely appear on tests as "true/false" questions. Such classification makes this quality characteristic an *attribute* – a parameter that cannot be represented as a continuous variable. In quality control, the terms "conforming" and "nonconforming" are used to describe two possible outcomes of inspecting a product for its conformance to specifications (Montgomery 1997a). In our case, a correct answer to a question means the answer is "conforming", and an incorrect (or "I do not know") answer means the answer is "non-conforming."

The underlying assumption behind the MBKP is that the statistic computed from any of the "B&A" questions is based on binomial distribution. The process of interest in our case is the transfer of knowledge in a classroom. We assume that the process of knowledge transfer is stable, and that the probability that a student will be able to answer any particular "B&A" question incorrectly is p. Moreover, we assume that the probability of one student answering a "B&A" question incorrectly is independent from the probability of another student answering the same question incorrectly. Therefore, each student's answer to a particular "B&A" question is a realization of a Bernoulli trial with the parameter p. In a class with N students, if D is the number of students answering a question incorrectly, then D has a binomial distribution with parameters N and p (Levine et al. 1998):

$$P\{D = x\} = \binom{N}{x} p^{x}(1-p)^{N-x}, \quad x = 0, 1, \ldots, N. \tag{4.2}$$

The population fraction nonconforming p is the ratio of the total number of incorrect answers in a population to the total number of answers in the population. The "population" could be a single class lectured by an instructor during a semester or could be the total flow of students through the particular course over a number of years.

Normally, the true fraction nonconforming p is not known. We can collect a sample n from the population (answers to a single B&A question from an MBKP asked on a particular day), and compute the sample fraction nonconforming $\hat{p}$:

$$\hat{p} = \frac{D}{n} \tag{4.3}$$

From the sample fraction nonconforming $\hat{p}$, the population fraction nonconforming p is, therefore, estimated. The sample fraction nonconforming $\hat{p}$ is the statistic that can be plotted on a statistical process control chart (for a brief overview of a statistical control chart, see Appendix VII). A fraction non-conforming *p*-chart will have the following characteristics (Montgomery 1997a):

$$\begin{aligned} UCL &= \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \\ Center &= \bar{p} \\ LCL &= \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \end{aligned} \tag{4.4}$$

where $\bar{p}$ is the estimate of the unknown population fraction nonconforming p:

$$\bar{p} = \frac{\sum_{i=1}^{m} \hat{p}_i}{m} \tag{4.5}$$

In Equation (4.5), m is the number of samples taken from a population.

4.5.3. SPC charts

As can be seen from the data in Appendix VII, the sample size n_i (number of students responding to a "B&A" question), varied from the case to case. The sample size was influenced by the student participation and attendance on a particular lecture day. The statistical process control chart with the parameters specified in Equation (4.4) assumes a constant sample size. One way to address the problem of a variable sample size is to use a standardized control chart (Montgomery 1997a). This chart has a center line set at zero; upper and lower control limits that are set at plus and minus three standard deviations, respectively; and the sample statistic plotted on the chart is computed by using the following formula:

$$Z_i = \frac{p_i - p}{\sqrt{\frac{p(1-p)}{n_i}}}. \tag{4.6}$$

In Equation (4.6), p_i is the sample proportion (P_{IA} or P_{IBCA}), p is the process estimate of the proportion, and n_i is the sample size (number of students returning the i-th MBKP). A standardized attributes SPC chart has the advantage of providing an opportunity to plot several p statistics coming from different processes on the same chart, since the points are plotted not in original, but in standard deviation units. In our case, this feature was used to compare the performance of different instructors, and the performance of the same instructor in different semesters.

4.5.4. Discussion of results: Instructors A and B (Fall 2002)

The goal of comparing the performance of the students in two sections of the same course during the same semester was to see if the factors presented below actually produced any difference in the final estimates of the process parameter p, and if the patterns of performance for each B&A question were similar. The factors studied were scheduling and class composition:

- Instructor A's section was scheduled early in the morning, and Instructor's B in the evening.

- Instructor B's class consisted of about 25% percent graduate and 75% undergraduate students attending lectures, and Instructor A's class had undergraduate students only.

An analysis of the patterns of both the P_{IBCA} and P_{IA} charts (see Figures 4.2, and 4.3, respectively) illustrates that the process of knowledge transfer was not in the state of statistical control in either section of the course. The similarity in patterns suggests that for some course concepts, knowledge transfer was consistently worse, or consistently better, than for some others.

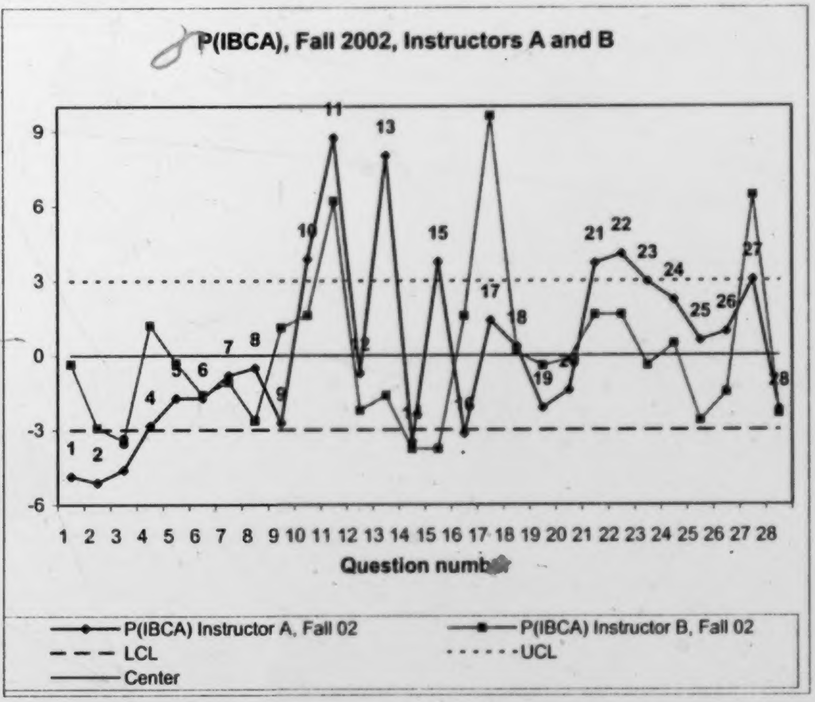


Figure 4.2: P_{IBCA} standardized attributes chart, Instructors A and B, Fall 2002

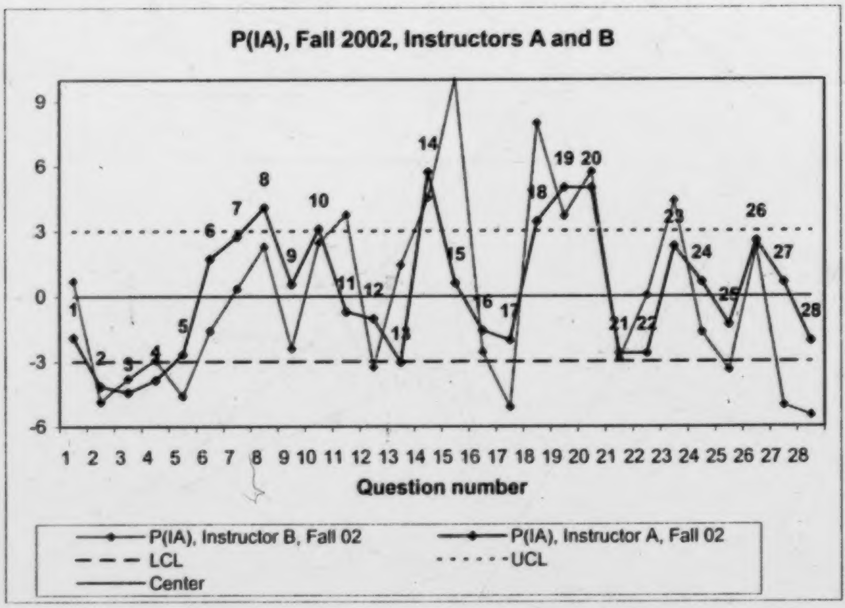


Figure 4.3: P_{IA} standardized attributes chart, Instructors A and B, Fall 2002

For example, questions 18, 19, and 20 were all related to the analysing of the cash flow statement and computing the operating cash flow. The high proportion of "Incorrect After" answers in both sections (see Figure 4.3) indicates that students did not learn this material well during the lecture when it was discussed.

Similarly, question 21 related to the calculation of operational management ratios. The high proportion of "Incorrect Before, Correct After" answers in Instructor A's section (see Figure 4.2) indicates that many students learned this concept during the lecture. The low proportion of "Incorrect After" answers in both sections (see Figure 4.3) indicates that after the lecture, relatively few students were unable to compute the required management ratios.

On some occasions, the students' relative performance was noticeably different, as points 13 and 15 on the P_{IBCA} chart illustrate (see Figure 4.2). The low proportion of "Incorrect Before, Correct After" answers in Instructor's B section resulted because the material related to question 13 was mentioned only briefly, and the material related to question 15 was not discussed at all.

The test for the equality of the two binomial parameters p was conducted to find if the difference in P_{IBCA} and P_{IA} statistics was significant between Instructor A and B. The statistics for testing a hypothesis were computed by using Equation (4.7) (Montgomery 1997a):

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \tag{4.7}$$

where

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}. \tag{4.8}$$

While the relative performance of students in both Fall 2002 sections was similar, the absolute performance was not. This finding is illustrated in Table 4.4:

Table 4.4: Absolute differences in performance (Instructors A and B, Fall 2002)

Statistic	Instructor A	Instructor B	Absolute difference	Weighted p	Z	Statistically significant?
P_{IBCA}	0.262	0.180	0.082	0.209	5.62	Yes
P_{IA}	0.195	0.311	-0.116	0.269	-7.29	Yes

The lower P_{IBCA} and higher P_{IA} for Instructor B's section indicate that students in this section learned less during the lectures than in Instructor A's section. Three factors may have contributed to this difference: two (different time of day, fraction of graduate students) were mentioned above, and the third factor is the lecturing styles. While the factor (or a combination of factors) that produced the difference cannot be determined with certainty (because factors may confound each other), the difference in lecturing styles may have contributed most to the difference in the overall P_{IA} and P_{IBCA} statistics:

- Instructor A designed the MBKPs and tried to cover all the subjects that were tested by a MBKP during a particular lecture. Instructor B did not see questions 9-24 before the lecture in which the MBKPs asking these questions were administered.
- While the two instructors agreed on what should be covered during a particular lecture, the relative importance of specific subjects differed for each instructor.
- Instructor B spent less time on the theory of a concept and more time on illustrating it with examples, while Instructor A did the opposite, spending more on theory, and less on the examples.

4.6.5. Discussion of results: Instructor A (Winter and Fall 2002)

Instructor A taught the same course during the Winter and Fall semesters in 2002. The same lecture notes and MBKPs were used during both semesters, except for the MBKP containing questions 9-12. As expected, the same lecturing style, the same MBKPs and the same lecture outlines produced very similar student performances, both in relative (see patterns in Figures 4.4 and 4.5), and in absolute terms (see Table 4.5).

Table 4.5: Instructor A, absolute differences in performance (Winter and Fall 2002)

Statistic	Fall 2002	Winter 2002	Absolute difference	Weighted p	Z	Statistically significant?
P_{IBCA}	0.262	0.276	-0.014	0.269	-0.87	No
P_{IA}	0.195	0.189	0.006	0.192	0.39	No

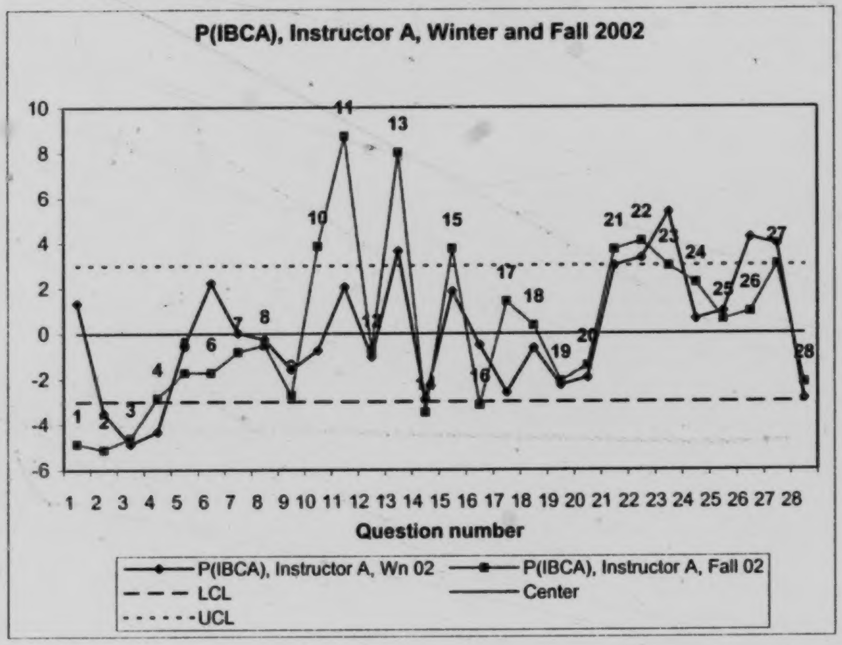


Figure 4.4: P_{IBCA} standardized attributes chart, Instructor A, Winter and Fall 2002

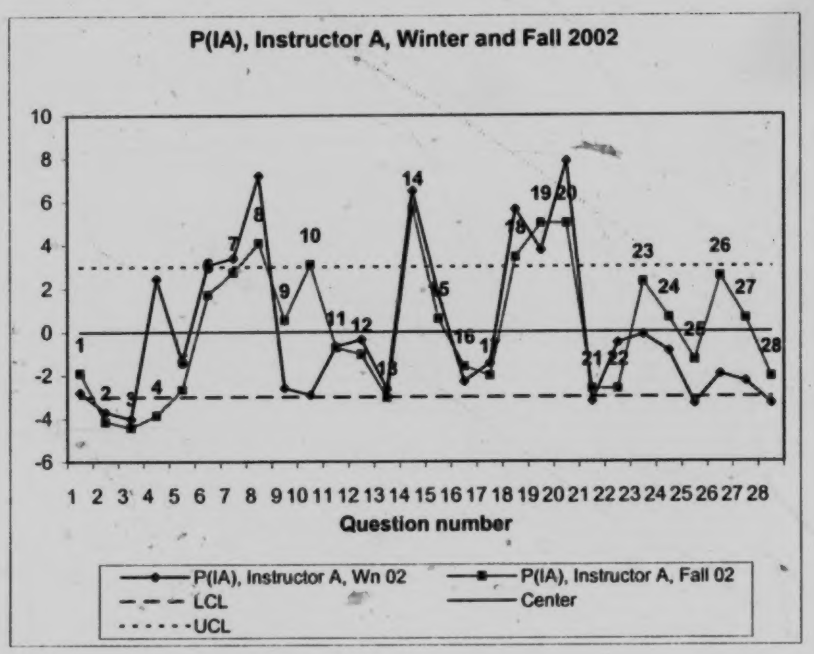


Figure 4.5: P_{IA} standardized attributes chart, Instructor A, Winter and Fall 2002

4.5.6. Discussion of results: Instructor A (Fall 2002 and Winter 2003)

Instructor A taught another section of the same course in the Winter 2003 semester. This time, though, Instructor C designed the B&A questions and administered the MBKPs. The following B&A questions from the Fall 2002 MBKPs were modified: 3, 4, 8, 10, 13, 14, 15, 16, 20, 23, 25, 26, 27, and 28. The course content and notes remained as they had been in Fall 2002. Instructor A provided Instructor C with a list of the knowledge blocks covered during those lectures when the MBKPs were administered, but Instructor A did not see the B&A questions themselves beforehand. This procedure was followed to prevent the instructor from shaping the lectures to provide answers to the questions posed (i.e., tutoring rather than teaching). The effect of these changes on the process of knowledge transfer is illustrated in Figures 4.6 and 4.7.

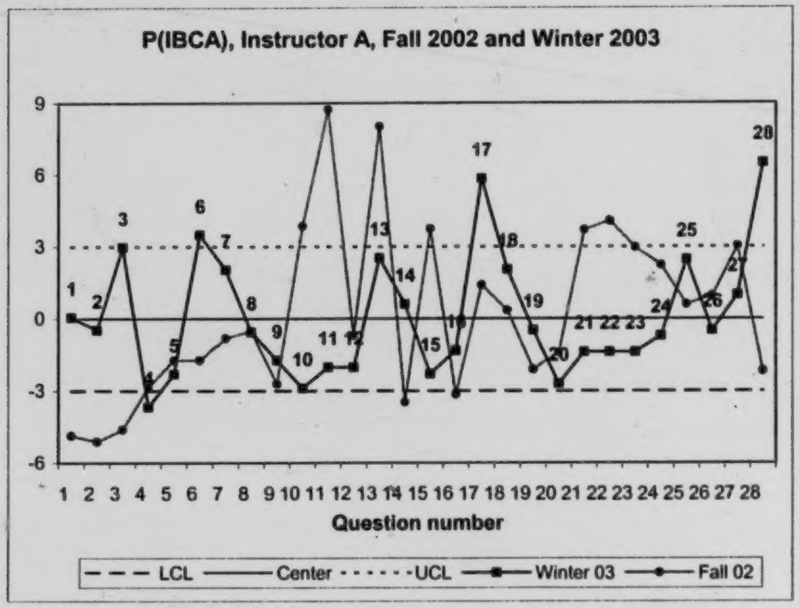


Figure 4.6: P_{IBCA} standardized attributes chart, Instructor A, Fall 2002 and Winter 2003

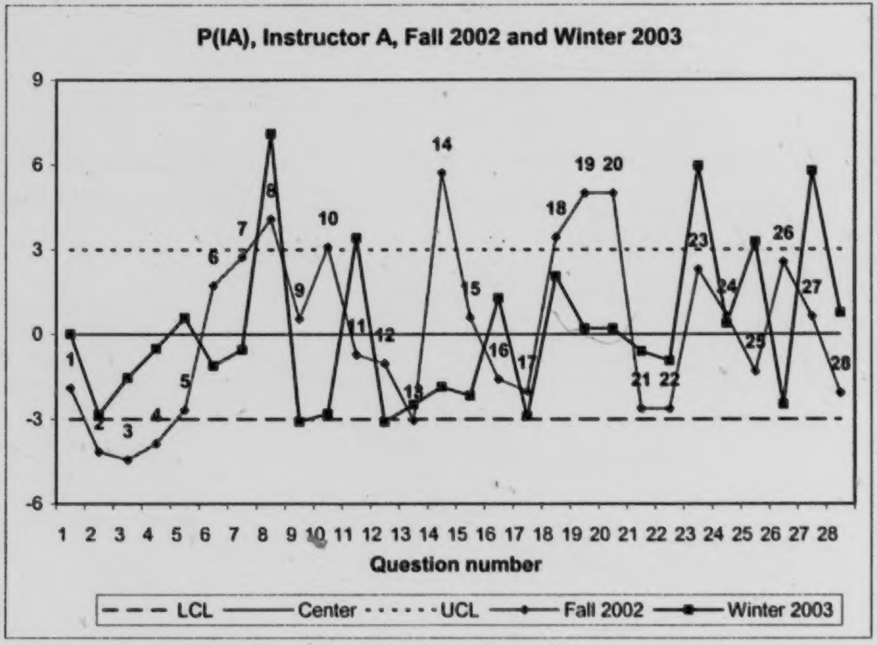


Figure 4.7: P_{IA} standardized attributes chart, Instructor A, Fall 2002 and Spring 2003

Some observations can be made about the process of knowledge transfer in the Winter 2003 semester. Firstly, this process still remained out of statistical control. This finding indicates that either other assignable causes produced out-of-control situations or corrective action (changing questions and lectures) failed to bring the process under statistical control.

Specific points indicate some remarkable changes. Questions 18, 19, and 20 were related to the cash flow statement – the subject that caused a high proportion of "incorrect after" answers in the Fall 2002 in both Instructor A's and Instructor B's classes (see Figure 4.3). During the Spring 2003 class, questions 18 and 19 were modified, and question 20 was left the same. The chart in Figure 4.7 indicates that during the Winter 2003 semester, the subject of cash flow statement did not cause many problems for the students.

Question 14 was related to the subject of cash versus book break-even. The low P_{IBCA} and high P_{IA} for question 14 during Fall 2002 indicated, on first sight, that few students had mastered that subject in the class, and that many had not known the correct answer to the question after the class. However, the question, rather than the material, appeared to be confusing students. Question 14 used the construct "not doing A is incorrect." During the Spring 2003 semester, the construct was changed to "doing A is correct". The higher P_{IBCA} and lower P_{IA}, compared to the corresponding statistics in Fall 2002, indicate that the question, and not the subject or lecture, was the likely assignable cause of out-of-control situation.

Analysis of relative performance produced the following results (see Table 4.6).

Table 4.6: Instructor A, absolute differences in performance (Fall 2002 and Winter 2003)

Statistic	Fall 2002	Winter 2003	Absolute difference	Weighted p	Z	Statistically significant?
P_{IBCA}	0.262	0.202	-0.060	0.229	-4.10	Yes
P_{IA}	0.195	0.215	0.020	0.206	1.43	No

The results show that statistically lower knowledge transfer occurred in Winter 2003 compared to that in Fall 2002 (as indicated by P_{IBCA}), while the proportion of "incorrect after" did not change significantly. The direction of the change indicates that the changes in MBKP design and administration had a negative effect on the process of knowledge transfer. Alternatively, it can be argued that changes in MBKP design and administration provided a closer estimate of the true population P_{IBCA} and P_{IA} parameters.

4.5.7. Autocorrelation considerations

If consecutive samples are even slightly correlated over time, the number of false alarms for a control chart can be significantly inflated, for control charts are sensitive to even small violations of independence (Montgomery 1997a, Quesenberry 1991a).

The nature of the educational process is such that subsequent knowledge is built on the foundation of the previous knowledge. In this course, for example, the subjects of double-entry and balance sheet equation are taught before students are introduced to a balance sheet and its construction by using the principles of double-entry and a balance sheet equation. Obviously, therefore, a student with a better knowledge of the double-entry and balance sheet equation will have a better understanding of the balance sheet.

In regard to the B&A questions, by knowing the answer to a double-entry B&A question, for example, a student might be in a better position to answer a balance sheet B&A question. This interdependence may introduce autocorrelation into the data.

This problem was addressed in two ways. Firstly, the B&A questions were designed so that any two, and, in particular, any two consecutive questions, could be answered independently: a student should be able to answer question *n* without knowing the answer to question *n-1*, and knowing the answer to question *n-1* should not increase the probability of answering question *n* correctly. Independence between successive questions is therefore achieved through the question design. To achieve independence of each answer within a sample, the students were instructed not to consult with each other while answering the questions.

Secondly, a sample autocorrelation function (Montgomery 1997a) was used to estimate the level of autocorrelation analytically. Under closer examination, we can recognize this function as the Durbin-Watson test, which measures the correlation of each residual (between the observation and process average) and the residual of a preceding observation. The details of the Durbin-Watson test can be found in Levine, Berenson, and Stephan (1998). The test results suggested that the null hypothesis (no autocorrelation) should not be rejected when using the level of significance $\alpha = 0.01$ (see Table 4.7).

Table 4.7: Results of the Durbin-Watson autocorrelation tests

Semester	Data set	Durbin-Watson statistic	Conclusion	
			$\alpha = 0.05$	$\alpha = 0.01$
Fall 2002, Instructor A	IBCA	1.683	No autocorrelation	No autocorrelation
	IA	1.399	No conclusion	No autocorrelation
Fall 2002, Instructor B	IBCA	1.591	No autocorrelation	No autocorrelation
	IA	1.777	No autocorrelation	No autocorrelation
Spring 2002, Instructor A	IBCA	1.380	No conclusion	No autocorrelation
	IA	1.563	No autocorrelation	No autocorrelation
Spring 2003, Instructor A	IBCA	1.410	No conclusion	No autocorrelation
	IA	2.444	No autocorrelation	No autocorrelation

4.5.8. Warm-up instability

The instability of a process during the "infancy" period of knowledge transfer is another possible issue. When a process is not stable at the beginning of the operation, all points obtained during this period may need to be ignored when setting the eventual control limits (Montgomery 1997a, Quesenberry 1991a). Warm-up instability may be present in data sets 1, 2, and 3 (see Figures 4.2, 4.3, 4.4, and 4.5) (the low P_{IA} combined with low P_{IBCA} values means that P_{CBCA} values were high for questions 1-4). This situation was expected. Questions pertaining to the first chapter of the class notes, which were discussed at the beginning of the class, had a very low percentage of incorrect answers, probably because the topics covered in the beginning were of a general nature, and answers to the questions could have been derived intuitively. Indeed, how hard is it for a student to answer the question "Is money a good measure of social value?"

The points collected during this "infancy" period (questions 1-4 in Figures 4.2, 4.3, 4.4, and 4.5) may have to be ignored during the analysis, because the statistics collected during that period do not provide a good estimate of the true population parameter.

In the future, when analysis is done in the on-going mode, practitioners should be aware of this situation and may want to administer questionnaires with questions that can be answered only with the knowledge obtained during a lecture and not with the knowledge brought from the outside.

The standardized P_{IBCA} chart based on the data sets 2 and 3 (Fall 2002, Instructors A and B), with the first four points ignored, is presented in Figure 4.8. The pattern of points did not change (compared to Figure 4.2), but some points that plotted outside the control limits on the original chart now plot inside, specifically the answers to questions 10, 21, 22 and 23. With the first four questions discarded, the average P_{IBCA} increased from 0.262 to 0.322 for Instructor A and from 0.180 to 0.189 for Instructor B.

Note that in the data set 4 (Instructor A, Spring 2003), the situation was different. The modified questions 3 and 4 produced a higher number of "incorrect after" answers (Figure 4.8), and, for question 3, the knowledge gain during a class was significantly higher than in Fall 2002 (Figure 4.6). We can argue that in Spring 2003, the warm-up instability was alleviated through the B&A questions' design.

While the SPC chart might suggest that no knowledge transfer (or low knowledge transfer) occurred during the lectures covered by questions 1 to 4 (as indicated on the P_{IBCA} chart), one must remember that the MBKP only measures the knowledge of facts gained over the last lecture period. Therefore, low P_{IBCA} does not imply that the instructor did not contribute to other aspects of students' knowledge (such as, for example, knowledge of experience, that will be recalled by students ten years from today). As a butterfly flopping its wings might start a hurricane across the globe, the instructor's influence on students' knowledge may not manifest itself in measurable ways immediately.

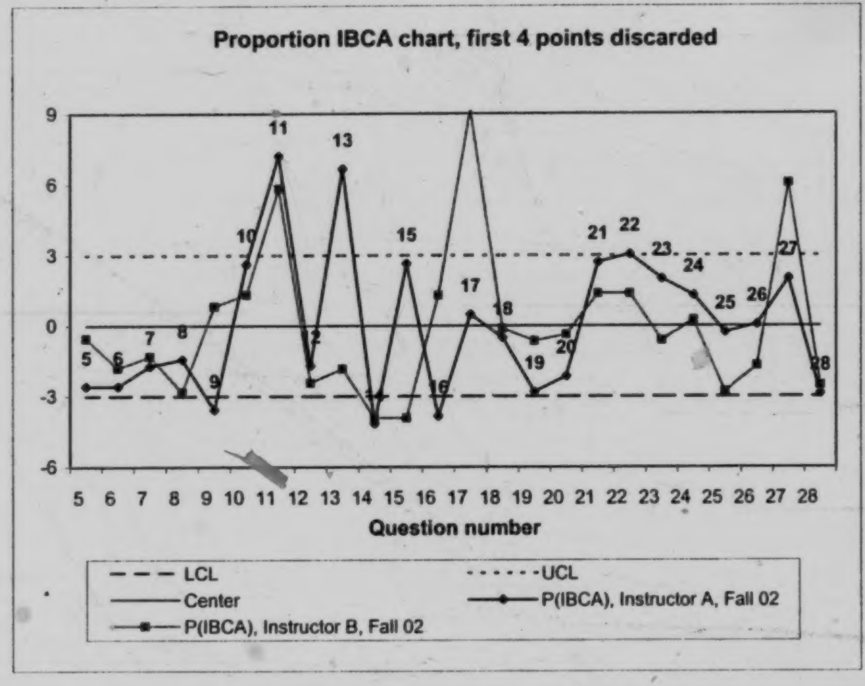


Figure 4.8: P_{IBCA} standardized attributes chart, first four points discarded, Fall 2002

4.5.9. Short-run process

SPC theory recommends collecting at least 25-30 samples for a variables chart, and at least 20-25 samples for an attributes chart, in order to establish reliable control limits (Hillier 1969, Montgomery 1997a).

The application of SPC in a classroom is a typical short-run problem: with a usual 13-week semester, about 25 samples are all an instructor is going to have. If the assignable causes can be detected for the points that plot outside the trial control limits, the samples producing those points should be discarded (Montgomery 1997a), thus leaving even fewer samples to work with. Another problem is that the reliable (but still questionable, if we have only 18-20 samples overall) control limits can be established only after the process p estimates have been obtained.

Estimates of P_{IA} and P_{IBCA} will be the most reliable after all samples have been collected, but they will not have been collected until the end of the semester. This situation makes statistical evaluation of the teaching/learning quality retrospective, rather then on-going (although using the obtained average and control limits in teaching the same course in the future might be reasonable if the course content and student body characteristics do not change).

Several authors discussed ways of dealing with the problem of short runs. Montgomery (1997) suggested using a standardized control chart for attributes, but when a process average is not known (as is usually the case for a new process) we still need an estimate of the process average. We cannot assume that we will know P_{IA} and P_{IBCA} beforehand. While an instructor implementing the proposed method may know from experience an average final grade in her/his class, whether anyone will be able to predict the average proportion of "incorrect before / correct after" answers when the proposed technique is introduced for the first time is still doubtful.

Hillier (1969) was one of the first researchers who worked on the problem of short runs. He offered a method for constructing variables short-run control charts, but did not explore the use of the approach for the attribute data. Quesenberry (1991b) proposed a method based on non-linear standardization transformation for short-run attribute data that allows starting charting with a second sample when the binomial parameter p is not known. When the parameter p is not known, the statistic Q is obtained as an inverse of the standard normal distribution with the argument being the cumulative hypergeometric distribution function. This function is an unbiased estimator of the binomial distribution function, and it converges to the binomial distribution as the number of the samples increases.

Figure 4.9 presents a *Q* chart obtained by using the procedure specified in Quesenberry (1991b). Since the analysis was done retrospectively, when all the data samples were already collected and the process average *p* calculated, the cumulative binomial distribution (with *p* values equal to P_{IBCA} or P_{IA}) was used to compute the *Q* statistic:

$$u_i = B(x_i, n_i, p) \qquad Q_i = \Phi^{-1}(u_i), \tag{4.9}$$

where:

B = cumulative binomial distribution with parameters *x* (number of occurrences in a sample), *n* (sample size), and *p* (probability of occurrence);
Φ^{-1} = inverse of the standard normal cumulative distribution (i.e., the *Z* value. For example, $\Phi^{-1}(0.05) = -1.64$).

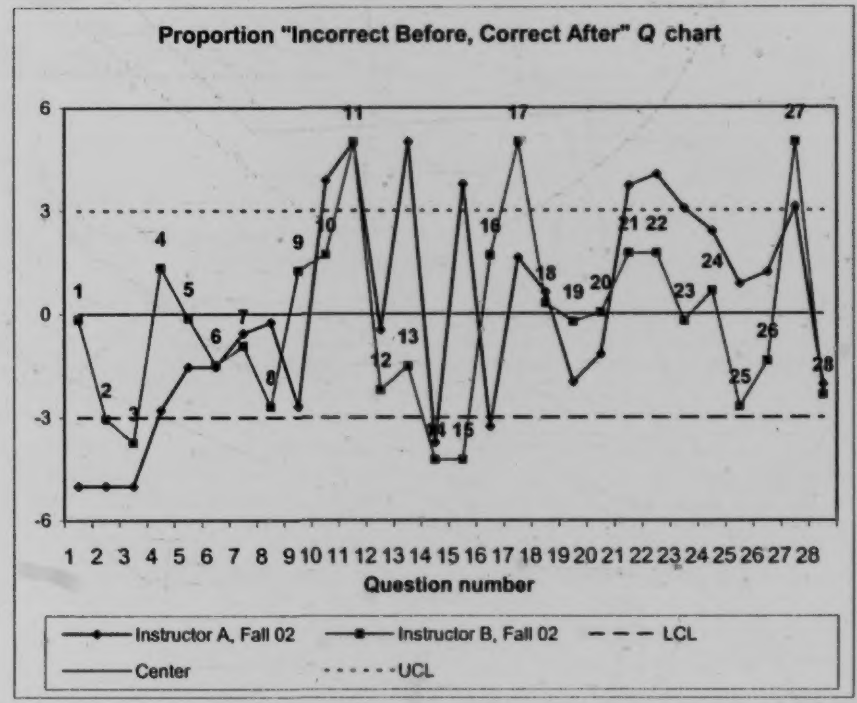


Figure 4.9: P_{IBCA} Q chart, Fall 2002

The *Q* chart, similar to the standardized *p* chart, has a center line at zero, and upper and lower control limits at plus/minus 3, correspondingly. The patterns of both P_{IBCA} charts (see Figures 4.2 and 4.9) are the same.

If a *Q* chart is to be used from the beginning of data collection, the cumulative hypergeometric distribution should be used to compute the *Q* statistic.

A *Q* chart should be used in the following cases:

- The total expected number of data samples is less than 20-25;
- The process parameters (mean, standard deviation) are not known;
- The establishment of a control chart is desired at the start of data collection.

4.6. Attributes vs. variables SPC chart: the problem of a small class

When a choice between the variables and attributes control charts is available, the variables chart is recommended. It provides a clearer picture about the process performance, indicating an approaching problem, and allowing for corrective action to be taken before the problem occurs (Montgomery 1997a).

An attributes control chart has its own set of advantages. The biggest one, probably, is measuring a student's answer to a question on a "correct – incorrect" scale. Measuring knowledge on a continuous scale, such as a score between 0 and 100, is difficult. Designing the 0-100 marking schemes is subjective and time-consuming. Normally, the problem is alleviated by including a number of smaller problems, each measured on its own "correct-incorrect" scale, and summing up the individual problem scores to obtain the grand total. This method, though, does not make the quality characteristic (knowledge) continuous – it only creates the attributes data at the lower level.

B&A questions that could be marked on a continuous scale have also to be longer and more complex, to give an evaluator a basis for assigning a fractional score. Students will require a longer time to answer such questions; therefore, fewer B&A questions can be asked. A typical semester has about 13 weeks, with some classes required for the midterm and final exams. This schedule leaves about 11 weeks of lectures, with the maximum number of lectures available for administering an MBKP being around 30 (if lectures are scheduled three times per week). An attribute assessment tool can be administered in every lecture, thus providing the aproximately 25-30 samples necessary to establish reliable control limits (Montgomery 1997a). If a variable control characteristic is to be collected, the number of points will be smaller, since a more time-consuming tool can be administered less frequently.

A significant drawback of the attributes chart is that it requires a large sample size to ensure a positive lower control limit. From Equation (4.4),

$$LCL = \overline{p} - 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}.$$

To ensure that LCL > 0, the sample size *n* has to be (Montgomery 1997a)

$$n > 9 \times \frac{1-\overline{p}}{\overline{p}}. \tag{4.10}$$

With the *p* values seen in this research being around 0.2 (P_{IBCA}, P_{IA}), the required sample size must be

$n > 9 \times (1 - 0.2) / 0.2 = 36.$

If the p value is lower (for example, if a goal for P_{IA} is zero), the sample size has to be even bigger. For example, for $p = 0.1$ to ensure LCL > 0, the n has to be greater than 81. Some classes (mostly graduate) have an enrolment limited to 20 students maximum, and some graduate classes have even lower enrolments. Also, frequently the class attendance is less than 100%. The fact that some students do not participate in the MBKP process further reduces the sample size.

If an MBKP is to be applied in a small class, the statistic collected has to be a variable.

4.7. MBKP with the statistic as variable: application

An opportunity to explore the applicability of a MBKP tool with the indicator of knowledge transfer that is a variable statistic presented itself in Spring 2003 in a form of a graduate Quality Management class taught by Instructor A, with an enrolment of 40 students. A three-hour long lecture was conducted once a week, for 13 weeks.

Sixteen B&A questions were designed by Instructor C. These questions are presented in Appendix IX. The answer to each B&A question was marked on a scale from 0 to 10. The statistic computed from a B&A question was the difference between a mark given for the answer after the lecture, and the mark given for the answer before the lecture. This statistic, named "knowledge gain" was, therefore, bound by interval [0, 10], with 0 corresponding to no knowledge gain, and 10 corresponding to the maximum knowledge gain.

An example of a MBKP measuring knowledge gain with two B&A question is presented in Figure 4.10.

ENGM XXX BEFORE AND AFTER QUESTIONNAIRE (Jan YY, 2003)

1. **In some credit card companies, departments issuing credit cards and departments collecting debt are managed and rewarded separately. The separation creates an incentive for the credit-issuing department to issue as many credit cards as possible, so their performance will look good. Credit cards are being issued to people with questionable credit history, which increases the amount of uncollectable debt. This increase worsens performance of the debt-collecting department. Briefly explain which quality principle is being overlooked in such a company?**

Before the lecture

After the lecture

2. **Name five differences between ISO 9000: 1994 series and ISO 9000: 2000 series**

Before the lecture

After the lecture

Figure 4.10: Example of a MBKP collecting a variable statistic

The knowledge gain data collected during the course are presented in Appendix X, and a data summary is given in Table 4.8 below:

Table 4.8: Data summary, variable statistic

Sample Number	Sample size	Sample Mean	Sample st dev			For X chart					For S chart		
				c4	A3	LCL	cent	UCL	B3	B4	LCL	cent	UCL
1	36	1.94	1.45	0.993	0.504	0.79	1.79	2.79	0.639	1.361	1.27	1.98	2.70
2	34	1.53	1.54	0.992	0.518	0.76	1.79	2.82	0.628	1.372	1.25	1.98	2.72
3	24	0.54	1.18		0.619	0.56	1.79	3.02	0.555	1.445	1.10	1.98	2.87
4	23	2.13	1.98		0.633	0.53	1.79	3.05	0.545	1.455	1.08	1.98	2.89
5	27	2.63	2.17	0.990	0.583	0.63	1.79	2.95	0.580	1.420	1.15	1.98	2.82
6	26	3.42	2.66	0.990	0.594	0.61	1.79	2.97	0.571	1.429	1.13	1.98	2.84
7	25	1.68	1.70		0.606	0.59	1.79	2.99	0.565	1.435	1.12	1.98	2.85
8	14	1.21	2.26		0.817	0.17	1.79	3.41	0.406	1.594	0.81	1.98	3.16
9	23	0.74	0.92		0.633	0.53	1.79	3.05	0.545	1.455	1.08	1.98	2.89
10	22	1.00	1.15		0.647	0.51	1.79	3.07	0.534	1.466	1.06	1.98	2.91
11	10	3.50	4.53		0.975	-0.1	1.79	3.73	0.284	1.716	0.56	1.98	3.41
12	10	3.70	4.06		0.975	-0.1	1.79	3.73	0.284	1.716	0.56	1.98	3.41
13	8	0.00	0.00		1.099	-0.4	1.79	3.97	0.185	1.815	0.37	1.98	3.60
14	13	1.54	1.05		0.850	0.10	1.79	3.48	0.382	1.618	0.76	1.98	3.21
15	9	1.33	2.18		1.032	-0.3	1.79	3.84	0.239	1.761	0.47	1.98	3.49
16	8	1.63	2.07		1.099	-0.4	1.79	3.97	0.185	1.815	0.37	1.98	3.60

The details for the construction of control limits can be found in Montgomery (1997a).

One can immediately notice a gradual decline in sample size from 36 to 8. This decline is typical in administration of MBKPs – as the air of novelty wore off, fewer students participated in the process. One of the remedies to declining interest and participation is to provide frequent feedback to students, in the form of an up-to-date SPC chart.

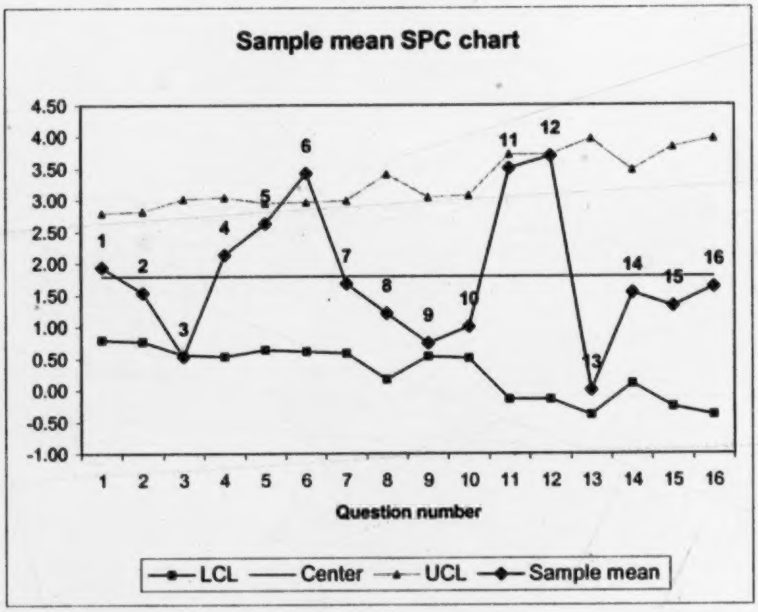


Figure 4.11: Sample mean SPC chart, variable characteristic

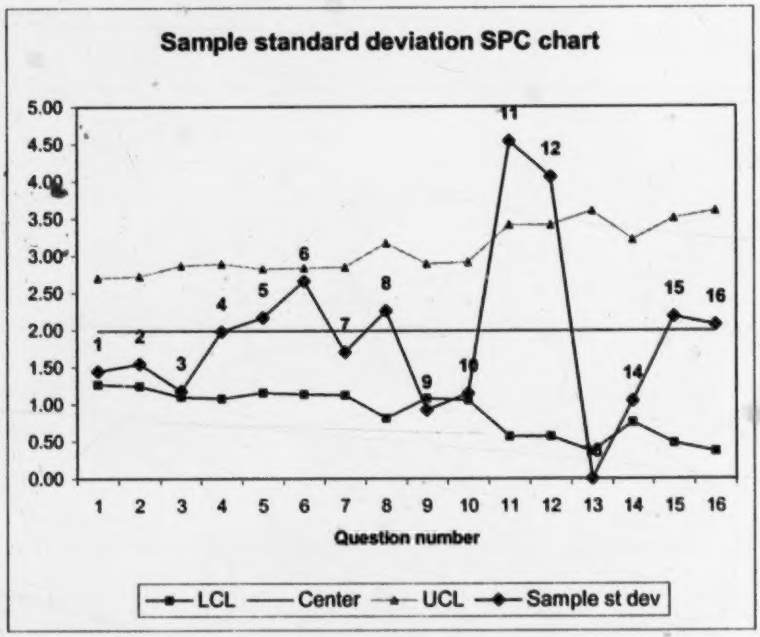


Figure 4.12: Sample standard deviation SPC chart, variable characteristic

Control charts for the sample mean and standard deviation are presented in Figures 4.11 and 4.12, respectively. The characteristic plotted on a sample mean chart is the sample average knowledge gain value, and on the standard deviation chart, the standard deviation of the sample.

Several observations can be made:

- The sample size consisting of 16 observations is not sufficient to establish reliable control limits. A short-run procedure may need to be applied. Procedures for designing short-run charts for variables can be found in Hillier (1969), Quesenberry (1991a), Quesenberry (1995), and Montgomery (1997a);
- The process of knowledge transfer, based on the computed control limits, was not in the state of statistical control;
- Due to a decrease in the sample size toward the end, the control limits become wider. Control limits are based on the sample standard deviation, which, in turn, is inversely proportional to a square root of the sample size;
- The lower control limit on the sample mean SPC charts falls below zero for samples 11-13 and 15-16. This finding by itself is not a problem, but, in this situation, there are no negative knowledge transfer values (which would mean students were losing knowledge during a lecture). No points can plot below the lower control limits in samples 11-13 and 15-16 (this situation is similar to that in Pierre and Mathios 1995). The lower control limit could be set at zero when calculations return a negative value.

Several points on the chart deserve special attention:

- Point 3: low mean, low standard deviation, sample means is below the lower control limit. The data indicate that very little knowledge gain occurred during the class. The analysis of the raw data and the question itself suggests that the question was of a general knowledge type, and the students were able to answer it well before the lecture (average of 5.17) and after the lecture (average of 5.71);

- Point 6: high mean, high standard deviation, sample mean is above the upper control limit. The data indicate that significant knowledge gain occurred during the class, but this knowledge gain was not uniform – some student learned a lot, while other learned little. The analysis of the raw data suggests that the subject was well-explained during the class ("average after" of 7.77), and that the students who did not know the subject before the class were able to learn it well. Those who knew the subject before, had little knowledge gain because of their good previous knowledge;
- Points 11 and 12: high mean, high standard deviation, standard deviation for both 11 and 12 is above the upper control limit. The analysis of the raw data and the questions indicated that the problem of quantification of knowledge manifested itself in this instance. Both questions 11 and 12 were designed to consist of two sub-questions worth 5 points each. In fact, the sub-questions were effectively marked on a "incorrect-correct" scale, with 0 for an incorrect answer, and 5 for a correct one. Therefore, the knowledge gain was measured as "all or nothing." The "average before" for question 11 was 1.20, and the "average after" was 4.70, while for question 12, the "average before" was 2.60, and the "average after" was 6.30. In a number of cases, individual answers received a mark of 1 before the class, and a mark of 10 after the class;
- Point 13: zero mean, zero standard deviation. The data suggest that no knowledge gain occurred at all uniformly for all students. The raw data, though, show an average of 3.38 before the class and an average of 3.38 after. After an analysis of the lecture, it was realized that the subject had not been covered during the lecture.

4.8. Search for assignable causes

As demonstrated in the previous examples, the application of SPC and MKBP in the evaluation of teaching and learning effects may produce a number of out-of-control points on the SPC chart. Figure 4.13 presents an algorithm that the instructor may use in seeking the assignable causes for each out-of-control situation found on the P_{IA} and P_{IBCA} charts.

For each point that plots outside of control limits, one must identify if an assignable or a random cause produced the out-of-control situation (Montgomery 1997a). Typically, two major reasons for poor performance will be suspected: poor lecture quality and poor question quality. In some cases, one can quite reasonably assume the underlying cause – the question (e.g., it had an error in it) or the lecture (if the subject of the question was not discussed at all, or if it was discussed in a hurry at the end of the lecture). In cases when the reason is not obvious, and to confirm a suspicion even if the reason seems to be clear, a designed experiment will be necessary to confirm or reject this hypothesis.

In most cases, while searching for the assignable causes of an out-of-control situation, a researcher will have to rely on the mental, rather that the experimental, analysis of a potential problem. While in industrial SPC applications conducting a statistically designed experiment to ascertain the assignable cause is often possible, in a social system such experimentation is often unethical and unfeasible. For example, if an instructor wanted to test the assumption of "lecture as the cause", a rigorous experimental design would require splitting the class in two groups, one used as the test group, and one as the control group. Obviously, if the lecture indeed was the cause of the out-of-control situation, the group receiving the "old", and assumingly poorer lecture, would be at a disadvantage.

Therefore, a researcher frequently will have to rely on intellectual deductions, rather than on statistical tests, in order to decide what produced the out-of-control situation.

If the instructor suspects question (Q) was the cause, the algorithm in Figure 16 can be used to test the assumption (the same algorithm can be used to test the lecture (L) quality as well; in this case, each Q should be replaced with L, and each L with Q in the algorithm). The instructor should start with the first point on the chart, and if this point is within the control limits, he or she would proceed to the next point (block "Next Point"). If a point is outside of the control limits, the instructor should consider whether a poor lecture or question quality might have been the cause. If the instructor believes neither one was the cause, then he or she will have to assume that the out-of-control situation was caused by random variation. Another possibility is that process of knowledge transfer is not truly stable, and for different topics, a difference in learning exists. If either question or lecture quality is suspected as a cause, the instructor should proceed as outlined in the algorithm.

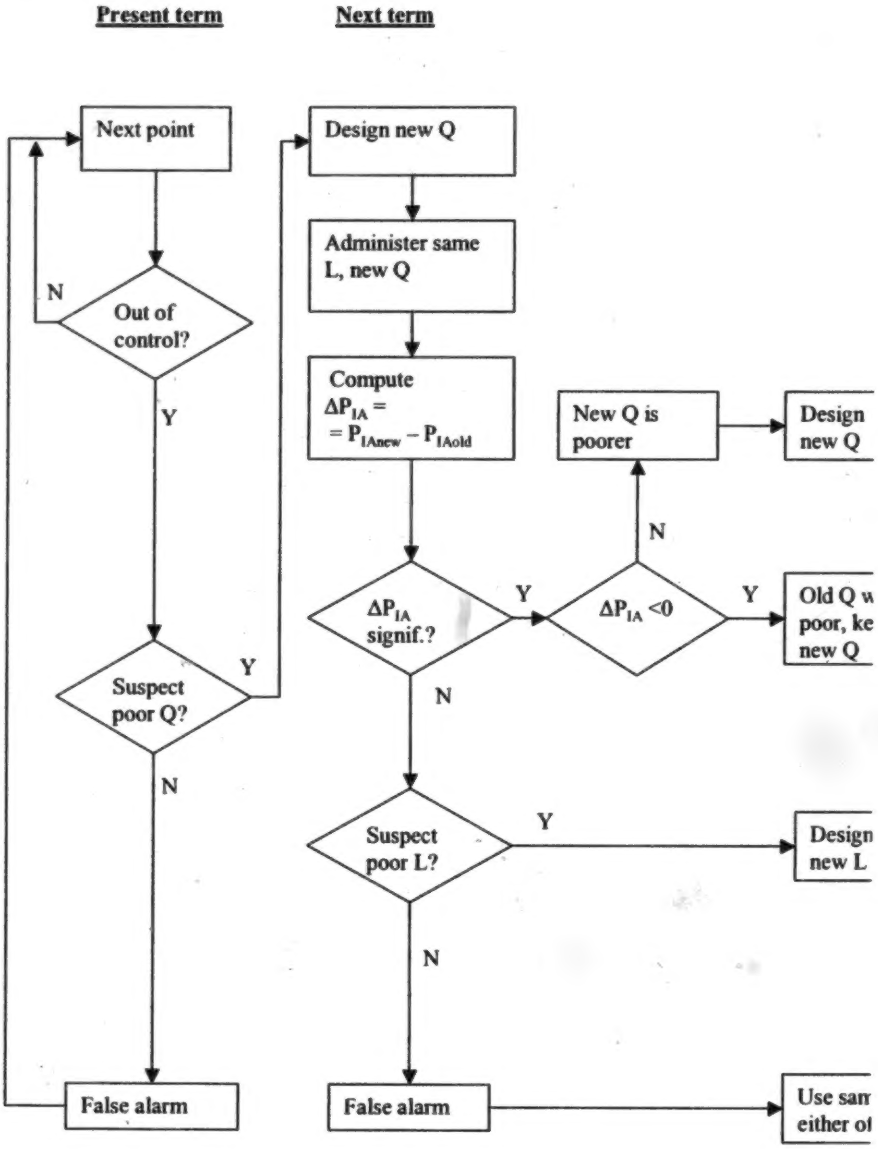

Present term
Next term
Next point
Out of control?
N
Y
Suspect poor Q?
Y
N
False alarm
Design new Q
Administer same L, new Q
Compute $\Delta P_{IA} =$ $= P_{IAnew} - P_{IAold}$
ΔP_{IA} signif.?
Y
N
$\Delta P_{IA} < 0$
N
Y
New Q is poorer
Design new Q
Old Q w poor, ke new Q
Suspect poor L?
Y
N
Design new L
False alarm
Use san either of

Figure 4.13: Algorithm for searching for an assignable cause

4.9. Suggestions for the Design of "Before and After" Questions

In order to reduce the ambiguity in data analysis, which is present when it is unclear if a poor question or a poor lecture produced the out-of-control situation, the following suggestions are given:

- Avoid "B&A" questions with differing degrees of difficulty. Their difficulty should be consistent throughout the course and within each MBKP. For example, one question asked, "Overestimating the depreciation period leads to writedowns – true or false?" To arrive at the correct conclusion, a student had to go through the following logical chain:
  a. "if A, then B" – an overestimated depreciation period causes a lower annual depreciation charge,
  b. "if B, then C" – a lower annual charge produces a higher remaining book value,
  c. "if C, then D" – a remaining book value higher then the market value produces a loss on a sale, which is recorded as a writedown,

  when normally, a logical chain "if A, then B" was being used. Questions should be simple, yet revealing, since they are not intended for student evaluation, but to measure student learning.
- Avoid atypical question constructs. For example, a high proportion of incorrect answers after the lecture to one of the questions may have been caused by the construct "not doing X is incorrect," while a better question construct would be "doing X is correct."
- Avoid unfamiliar terms. A high P_{IA} proportion to one question was probably due to confusion when the term "depreciation" (used consistently in the course) was replaced with the term "amortization" (mentioned only once at the beginning of the course).
- Avoid self-design of "B&A" questions. In other words, an instructor should consider delegating the design of questions to someone else, for example,

another instructor who taught the same course before. The instructor may discuss with the teaching assistant the question in general, but should not see the question before it is administered. This procedure should prevent the instructor from perhaps inadvertently concentrating on providing the answers to the B&A questions.

- When measuring knowledge transfer on a continuous, rather than "incorrect-correct" scale, also avoid breaking a single B&A question into sub-questions marked on a "incorrect-correct" scale. Design questions so that the gradual knowledge gain can be captured.

4.10. A designed experiment as a tool to discover assignable causes

For most data points falling outside the trial control limits on the SPC chart, an assignable cause may be suspected. For example, if the question was not covered during the lecture, we would have a high P_{IA} value. For some other points, however, the assignable causes cannot be identified easily, since several factors might confound each other in producing the effect. Was the factor a poor question, a poor lecture, or a combination of both?

A standard feature of the course, which was the subject of this study, was the collection of students' feedback at the end of the course. This feedback was incorporated into the design of subsequent lectures, notes, assignments, and exams. Therefore, when we look at the relative performance of the same instructor on the same lecture over time, we have to take into account that besides natural variation in the quality of lecturing, a conscious quality improvement resulted from an increase in the instructor's experience, knowledge, and effort.

The only way to resolve confounding and to confirm a suspicion even when the real cause seems to be obvious is to conduct a designed experiment. Indeed, Angelo and

Cross (1993) mention: "Through…the design of modest classroom experiments, teachers can learn … how students respond to particular teaching approaches" (p. 5).

The strategy of investigating whether a poor lecture or question quality produced an out-of-control situation was described in section 4.8. and presented in the algorithm in Figure 4.13. The methods of designing statistical experiments and computing ΔP_{IA} and ΔP_{IBCA} are presented in Appendix XI.

4.11. Limitation

One must realize that "knowledge" is an extremely broad domain. Eftekhar (1998) distinguished between the knowledge gained from personal experience, and the knowledge acquired through instruction. We may further break down the concept of knowledge into the "matter" (i.e., theory, facts, applications), the "source" (i.e., self-education, personal experience, observation, instruction), and the "timeframe" (i.e., gained over one hour, one semester, or the lifetime).

The Modified Background Knowledge Probe may be used to test the knowledge of either theory, facts, or applications, but is intended to measure only the knowledge gained through instruction (i.e., knowledge transfer from the instructor) over a relatively short period of time (one lecture).

4.12. Summary

Chapter 4 described how the process of knowledge transfer can be monitored by using one of the tools of statistical quality control – the statistical process control chart. The process statistics were collected by using a Modified Background

Knowledge Probe. The combination of the MBKP and the SPC allowed for the identification of lectures during which a low knowledge transfer occurred and for addressing the problem shortly thereafter. The performance of two different instructors teaching two sections of the same engineering management course was compared both in relative and absolute terms. The performance of the same instructor teaching the same course in different time was illustrated as well.

While originally developed to measure knowledge transfer by collecting attribute statistics in large classes, the MBKP tool was modified to collect variable statistics in classes with a low enrollment.

For instructors interested in introducing the MBKP tool into their classes, the implementation algorithm and suggestions for the design of Before and After questions were provided.

Chapter 5: Discovering Cause-And-Effect Relationships

5.1. Introduction

It was argued in Chapter 3 that the process of knowledge transfer from the instructor to students is at least partially, responsible for such educational outcomes as the students' test results, which themselves are indicators of the students' overall knowledge. Students can also gain knowledge from self- and peer-education. Many other factors, such as family background (Haveman and Wolfe 1995), language and ethnic background (Worthington 2002 ,Stiefel et al. 2001), gender and age (Worthington 2002), attitude toward the subject (George and Kaplan 1998) and toward the instructor (Murray et al. 1990, Cranton and Smith 1990), and level of enrollment (Worthington 2002), n[illegible]nly may contribute to the students' academic performance, but may influence each other in a complex cause-and-effect relationship. In this chapter, a SEM model will be created to discover the cause-and-effect relationships among the variables of a classroom system.

5.2. Model variables and hypothesized relationships

In order to determine how exactly student performance is being influenced in a classroom, an instructor might appear to have to collect a significant amount of data on many variables. Performance measurement theory and practice, however, calls for the selection and monitoring of a "vital few" indicators in order not to become overburdened with data and data-collection efforts (Andersen and Fagerhaug 2002, Hatry 1999, Wholey 1999, Smith 1995).

A set of performance indicators, arranged in a framework, was proposed in Section 3.5. It was decided to test whether a model of classroom performance could be created based on the data provided by the proposed educational performance framework. A structural model will be created based on a number of hypothesized casual relationships among variables in a classroom educational system. The model's latent variables and their indicators, with corresponding Greek notation (see Appendix II for details), are presented in Table 5.1.

A number of causal relationships among the model's latent concepts were postulated *a priori*. These relationships are described in Table 5.2.

Table 5.1: Variables and indicators in the SEM model of classroom performance

Theoretical concept	SEM Variable label	Observed indicator	SEM Observed indicator & label
Financial background	ξ_1	Report on owning a personal computer	OWNACOMP x_1
Gender	ξ_2	Reported gender	STUDTSEX x_2
Importance of having fun while in university	ξ_3	Reported importance of having fun in university	HAVEFUNU x_3
Importance of succeeding academically	ξ_4	Reported importance of doing well in university	DOWELLUN x_4
Language background	ξ_5	Reported language background	LANGBGND x_5
Age	ξ_6	Reported age	STUDTAGE x_6
Enrollment level	ξ_7	Reported enrollment level in the course	GRADLEVL x_7
Instructor's teaching experience	ξ_8	Years teaching	INSTEXPR x_8
Extra-curricular activities	η_1	Reported participation in sports and cultural events	EXTRACTV y_1
Language practice	η_2	Report on speaking English in everyday life	SPEKENGL y_2
Academic background in the discipline	η_3	Report on taking a similar course previously	SMLRCRSE y_3
Time devoted to self-studying	η_4	Reported homework time	HWRKTIME y_4
		Reported time spent reading text/notes	READTEXT y_5
Attendance	η_5	Reported number of lectures missed	LTRSMISS y_6
Student knowledge	η_6	Midterm test score	MIDTSCRE y_7
Perceived course workload	η_7	Reported perceived workload	WORKLOAD y_8
Satisfaction with the course in general	η_8	Reported satisfaction with the course	SATISFCT y_9
Satisfaction with the instructor	η_9	Reported satisfaction with the instructor	INTRGJOB y_{10}
Attitude toward the course subject	η_{10}	Reported attitude toward subject	LIKESBJT y_{11}
Balancing academic performance and recreation while at university	η_{11}	Report on possibility of doing well and having fun while at university	DOWLHVFN y_{12}

Table 5.2: Postulated causal relationships in the SEM model of classroom performance

Structural coefficient	**Hypothesis**	**Expected sign**
β_{41}	Students spending more time on extra-curricula activities will have less time for self-studying.	-
β_{61}	Attendance of concerts, art events, participation in sports are the signs of a "well-rounded" individual, who will score higher on a test (e.g.,"All work and no play makes Jack a dull boy").	+
β_{42}	Individuals who speak English less frequently will spend more time self-studying.	-
β_{62}	Non-speakers may not understand some questions or parts of them; also may not be able to express themselves as clearly as native speakers on questions that require written answers.	+
β_{63}	Students who took a similar course previously will know some material and will have an advantage on exams.	+
β_{73}	Students who took a similar course may be in a better position to evaluate the workload by comparing it to that in those similar courses.	±
β_{83}	Students who took a similar course will compare their satisfaction with the course against the satisfaction with similar courses.	±
β_{93}	Students who took a similar course will compare their satisfaction with the instructor against the instructors in similar courses.	±
β_{64}	Students who spend more time self-studying will have better knowledge of the subject material and will be better prepared for a test.	+
β_{74}	Students who spend more time self-studying will perceive that the workload is higher than that in other courses.	+
β_{45}	Students who miss more lectures will have to spend more time self-studying; it is also possible that decreased attendance also will discourage students from studying.	±
β_{65}	Students who missed more lectures will be less prepared for the tests, as they contain mostly questions discussed in class.	-
β_{46}	Students who received low marks on the test will spend more time self-studying after the test.	-
β_{86}	Students who receive a lower grade will be dissatisfied with the course, as they will believe that "course failed them."	+
β_{96}	Students who receive lower grade will be dissatisfied with the instructor, as they will perceive that instructor did not grade their abilities fairly.	+
β_{87}	Students who perceive that course workload is higher than typical will be less satisfied with the course.	-
β_{97}	Students who perceive that course workload is higher than typical will be less satisfied with the instructor, since the instructor sets the number and difficulty of homework assignments and exams.	-
β_{58}	Students who like the course will be less likely to miss lectures.	-
$\beta_{10\text{-}8}$	Students who are satisfied with the course will feel more positive about the course subject.	+
β_{59}	Students who like an instructor will miss fewer of the instructor's lectures	-

Table 5.2 continued

Structural coefficient	Hypothesis	Expected sign
β_{89}	Students who are satisfied with the instructor will be more likely to be satisfied with the course as well[1].	+
$\beta_{10\text{-}9}$	Students who like instructor will feel more positive about the course subject.	+
$\beta_{4\text{-}10}$	Students who like subject of the course might spend more time exploring additional aspects of the subject.	+
$\beta_{5\text{-}10}$	Students who like the subject of the course will miss fewer lectures.	-
$\beta_{1\text{-}11}$	Students who believe they can to do well and have time for fun will engage in more extra-curricula activities.	+
γ_{13}	Students who think having fun is important, will spend more time on out-of-class activities (e.g., on "fun").	+
γ_{73}	Students who think having time for fun will perceive course workload as being high and infringing on their free time.	+
$\gamma_{11\text{-}3}$	Students who think having time for fun is important, will have a stronger opinion on whether having time for fun and still doing well is possible; the relationship, though, could be of either sign.	±
γ_{44}	Students who understand the importance of doing well at university will be spending more time self-studying in order to achieve a higher grade.	+
γ_{54}	Students who place importance on doing well in university will miss fewer lectures, since lecture attendance is important for succeeding in this class.	-
$\gamma_{11\text{-}4}$	Students who think doing well in university is important, will feel they do not have enough time for doing well in school and having fun at the same time.	-
γ_{15}	Students who are non-native English speakers may not benefit from extra activities such as movies, theater; also, foreign students typically have less free time available.	-
γ_{25}	Foreign-born students will speak English less frequently.	-
γ_{45}	In a discussion-type class, non-native English speakers may not understand everything discussed in class, and therefore may have to spend more time on reading textbook/notes.	+
γ_{65}	Non-native speakers are at a disadvantage during the tests as they might not understand some texts, some concepts (e.g., taxes), and may not be able to express their thoughts clearly on written answers.	-
γ_{75}	Non-native English speakers have to make more of an effort during lecture discussions, assignments, and exams.	+

---

[1] While it was assumed that the sign of the structural coefficient β_{89} will be positive, the negative sign is also possible. From personal experience, one might think of the military basic training (a.k.a. "boot camp"), where the animosity of recruits toward the instructors is intentionally fostered in order to build teamwork among the recruits and increase their motivation. The recruit's goal becomes to demonstrate that she or he can succeed against the instructors' "ill will", and satisfaction with the course comes from one's accomplishments despite the (possible) dislike of a particular instructor. The realization that such scenario was carefully scripted usually comes later.

Table 5.2 continued

Structural coefficient	Hypothesis	Expected sign
γ_{16}	Older students have more responsibilities – family and jobs – and, therefore, have less free time available for extra-curricula activities.	-
γ_{46}	Older, more mature students have a more responsible attitude toward studying.	+
γ_{86}	Older students have a more mature approach to evaluating the quality of a course.	±
γ_{96}	Older students have a more mature approach to evaluating the quality of instruction.	±
γ_{17}	Graduate students are older, and therefore will have less time for extra activities; as well, they need to spend more time on research, which, again, leaves less free time.	-
γ_{37}	Graduate students have taken more courses than undergraduate students, and, therefore, have a greater chance of having taken a similar course.	+
γ_{47}	For graduate students, the importance of spending more time on homework should be more obvious from experience, as well from the need to maintain good marks for scholarship and other academic requirements.	+
γ_{57}	Graduate students have a more responsible attitude toward studying, and will miss fewer lectures.	-
γ_{67}	Graduate students may have taken more courses on related subjects and, therefore, will have better knowledge of the material; also, graduate students may be more experienced in test-taking.	+
γ_{87}	Graduate students have taken more courses in general and are in a better position to evaluate courses objectively.	±
γ_{97}	Graduate students have seen more instructors and are in a better position to distinguish between a good and a bad instructor.	±
γ_{10-7}	For graduate students, this is an elective course; only those who like it will choose it.	+
γ_{48}	Less experienced instructors may not have a well-balanced set of homework assignments, and students will spend either more or less time on homework than those students in the class of an experienced instructor.	±
γ_{58}	Less experienced instructor's lectures are less entertaining than those of a more experienced instructor.	-
γ_{88}	Students will be more satisfied with a course taught by a more experienced instructor.	+
γ_{98}	Students like experienced instructors.	+
γ_{10-8}	Instructors with greater teaching experience are better able to generate interest in a subject	+

5.3. Data

The data on student attitudes, perceptions, and performance in the course were collected by using a questionnaire that contained performance indicators described in Section 3.5. The questions were matched with the SEM variables, as indicated in Table 5.3.

Table 5.3: Student questionnaire and corresponding SEM variables

Question	Response categories	SEM concept	Observed indicator & label
What was your midterm test score?	1. below 60 2. 60 – 69.5 3. 70 – 79.5 4. 80 – 89.5 5. 90 – 100	η_6	MIDTSCRE y_7
Do you agree with the following statement: "I like the subject of this course"?	1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree	η_{10}	LIKESBJT y_{11}
Do you agree or disagree with the following statement: "The instructor is doing a good job in making this course interesting"?	1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree	η_9	INTRGJOB y_{10}
How much time (in hours per week) do you usually spend on homework for this course?	1. No time 2. Less then 1 hr 3. 1 – 3 hours 4. 3 – 6 hours 5. 6 – 8 hours 6. more than 8 hours	η_4	HWRKTIME y_4
How often do you use / read textbook (course notes)?	1. Never 2. Once in several weeks 3. Once a week 4. Several times per week 5. Every day	η_4	READTEXT y_5
How many lectures have you missed?	1. None 2. A few 3. About one a week 4. About two a week 5. Most of the lectures	η_5	LTRSMISS y_6
What do you think about the workload in this course?	1. Too easy 2. Easy 3. About average 4. Hard 5. Too hard	η_7	WORKLOAD y_8
Which best describes your satisfaction with the course:	1. Unsatisfied 2. Somewhat unsatisfied 3. Neutral 4. Rather satisfied 5. Satisfied	η_8	SATISFCT y_9
Do you think it is important to do well in a university?	1. No 2. Yes	ξ_4	DOWELLUN x_4
Do you think it is important to have time to have fun?	1. No 2. Yes	ξ_3	HAVEFUNU x_3

Table 5.3 continued

Question	Response categories	SEM concept	Observed indicator & label
Do you think it is possible to do well in a university and have time to have fun?	1. No 2. Yes	η_{11}	DOWLHVFN y_{12}
What is your age group?	1. 20 years or lower 2. 20 – 25 years 3. Over 25 years	ξ_6	STUDTAGE x_6
What is your gender?	1. Male 2. Female	ξ_2	STUDTSEX x_2
What is your language background?	1. English-speaking background domestic or overseas student 2. Non-English-speaking background domestic student 3. Non-English-speaking background overseas student	ξ_5	LANGBGND x_5
What is your level of enrollment in this class?	1. Undergraduate 2. Graduate	ξ_7	GRADLEVL x_7
Do you own a personal computer?	1. No 2. Yes	ξ_1	OWNACOMP x_1
Have you previously taken courses in a similar area (e.g., finance, accounting, management)?	1. No 2. One course 3. Two courses 4. Three or more courses	η_3	SMLRCRSE y_3
How often do you speak the language of instruction in everyday life?	1. Never 2. Sometimes 3. Often 4. Always or almost always	η_2	SPEKENGL y_2
During a typical semester, how often do you play sports, go to the movies, attend a concert, etc.?	1. Never 2. Rarely 3. About once a month 4. About once a week 5. About every day	η_1	EXTRACTV

The survey using the questionnaire was conducted in the undergraduate course in Engineering Management taught at the Department of Mechanical Engineering at the University of Alberta. Sometimes a section of the course is also available to the graduate students. The course is compulsory for all undergraduate engineering students, who have the option of choosing either a more calculus-oriented third-year course, or a more discussion-oriented fourth-year course. This course is an elective course for the graduate students.

The anonymous questionnaire was administered after the first midterm exam, in five sections over two semesters (Fall 2003 and Winter 2004). There were three undergraduate fourth-year courses, one undergraduate third-year course, and one fourth-year course with undergraduate and graduate students enrolled. The range of each response category in question 1 (midterm test score) was adjusted for each section to obtain an approximately equal distribution of low-, average-, and high-scoring students across all sections. The categories for the variable "Instructor's Teaching Experience" were "1", for the least teaching experience through "4" for the most teaching experience.

The students' participation in the research was anonymous and voluntary. The students were informed that their decision to participate or not to participate in the research would have no effect on their final mark in the course. The response rate to the questionnaire varied between 53% to 85% based on the number of students enrolled in the class (the actual response rate, based on the number of students present during the day the questionnaire was administered, would be higher). The total number of returned questionnaires was 396. After the list-wise deletion of missing data, the sample size became 384 questionnaires. Based on the questionnaires' data, a matrix of covariances among the observed indicators (matrix *S*) was created. This matrix is presented in Appendix XII.

5.4. LISREL model

The model will be described in the equation form this section, while the model in the path diagram form is presented in Appendix XIII. Specific model elements deserving special attention will be highlighted in pictorial form in this section.

The model, named "Original", has nineteen latent concepts (eight exogenous, ξ_1 - ξ_8, and eleven endogenous, $\eta_1 - \eta_{11}$). All concepts, except η_4 ("Time devoted to self-studying") have a single indicator. Concept η_4 is measured by two indicators, y_4 ("Reported homework time") and y_5 ("Reported time spent reading text/notes").

The structural relations among the variables in the model will be expressed by using equations (III.1)-(III.3) (see Appendix III). The resulting model specifications are presented in Equations (5.1) – (5.3).

The four remaining matrices required for model specification have the following parameters:

– Φ (matrix of covariances among exogenous concepts) is a symmetrical 8x8 matrix with freed coefficients to be estimated by the model;

– Ψ (matrix of covariances among ζ errors) is a diagonal 11x11 matrix with diagonal elements (representing error variances) freed to be estimated by the model, and non-diagonal elements fixed at zero, in accordance with the assumption that the errors are distributed independently;

– Θ_ε (matrix of covariances among errors ε) is a diagonal 12x12 matrix with diagonal elements (representing error variances for measuring endogenous variables) fixed at specified values (for reasons discussed in Section 5.4.2) and non-diagonal elements fixed at zero, in accordance with the assumption that the errors are distributed independently;

– Θ_δ (matrix of covariances among errors δ) is a diagonal 8x8 matrix with diagonal elements (representing error variances for measuring exogenous variables) fixed at specified values (for reasons discussed in Section 5.4.2) and non-diagonal elements fixed at zero, in accordance with the assumption that the errors are distributed independently.

From equation (III.1),

$$\boldsymbol{\eta} \quad = \qquad \mathbf{B} \qquad * \quad \boldsymbol{\eta}$$

$$\begin{bmatrix}\eta_1\\ \eta_2\\ \eta_3\\ \eta_4\\ \eta_5\\ \eta_6\\ \eta_7\\ \eta_8\\ \eta_9\\ \eta_{10}\\ \eta_{11}\end{bmatrix} = \begin{bmatrix} & & & & & & & & & & \beta_{1\_11}\\ & & & & & & & & & & \\ & & & & & & & & & & \\ \beta_{41} & \beta_{42} & & & \beta_{45} & \beta_{46} & & & & \beta_{4\_10} & \\ & & & & & & & \beta_{58} & \beta_{59} & \beta_{5\_10} & \\ \beta_{61} & \beta_{62} & \beta_{63} & \beta_{64} & \beta_{65} & & & & & & \\ & & \beta_{73} & \beta_{74} & & & & & & & \\ & & \beta_{83} & & & \beta_{86} & \beta_{87} & & \beta_{89} & & \\ & & \beta_{93} & & & \beta_{96} & \beta_{97} & & & & \\ & & & & & & & \beta_{10\_8} & \beta_{10\_9} & & \\ & & & & & & & & & & \end{bmatrix} * \begin{bmatrix}\eta_1\\ \eta_2\\ \eta_3\\ \eta_4\\ \eta_5\\ \eta_6\\ \eta_7\\ \eta_8\\ \eta_9\\ \eta_{10}\\ \eta_{11}\end{bmatrix} +$$

$$+ \qquad \boldsymbol{\Gamma} \qquad * \ \boldsymbol{\xi} + \boldsymbol{\zeta}$$

$$+ \begin{bmatrix} & & \gamma_{13} & & \gamma_{15} & \gamma_{16} & \gamma_{17} & \\ & & & & \gamma_{25} & & & \\ & & & & & & \gamma_{37} & \\ & & & \gamma_{44} & \gamma_{45} & \gamma_{46} & \gamma_{47} & \gamma_{48}\\ & & & \gamma_{54} & & & \gamma_{57} & \gamma_{58}\\ & & & & \gamma_{65} & & \gamma_{67} & \\ & & \gamma_{73} & & \gamma_{75} & & & \\ & & & & & \gamma_{86} & \gamma_{87} & \gamma_{88}\\ & & & & & \gamma_{96} & \gamma_{97} & \gamma_{98}\\ & & & & & & \gamma_{10\_7} & \gamma_{10\_8}\\ & & \gamma_{11\_3} & \gamma_{11\_4} & & & & \end{bmatrix} * \begin{bmatrix}\xi_1\\ \xi_2\\ \xi_3\\ \xi_4\\ \xi_5\\ \xi_6\\ \xi_7\\ \xi_8\end{bmatrix} + \zeta \qquad (5.1)$$

From equation (III.2),

$$\mathbf{y} \quad = \quad \boldsymbol{\Lambda}_{\mathbf{Y}} \quad * \quad \boldsymbol{\eta} \quad + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \begin{bmatrix} 1 &&&&&&&&&& \\ & 1 &&&&&&&&& \\ && 1 &&&&&&&& \\ &&& \lambda^{y}{}_{44} &&&&&&& \\ &&& 1 &&&&&&& \\ &&&& 1 &&&&&& \\ &&&&& 1 &&&&& \\ &&&&&& 1 &&&& \\ &&&&&&& 1 &&& \\ &&&&&&&& 1 && \\ &&&&&&&&& 1 & \\ &&&&&&&&&& 1 \end{bmatrix} * \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \eta_5 \\ \eta_6 \\ \eta_7 \\ \eta_8 \\ \eta_9 \\ \eta_{10} \\ \eta_{11} \end{bmatrix} + \varepsilon \qquad (5.2)$$

From equation (III.3),

$$\mathbf{x} \quad = \quad \boldsymbol{\Lambda}_{\mathbf{X}} \quad * \quad \boldsymbol{\xi} \quad + \boldsymbol{\delta}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix} = \begin{bmatrix} 1 &&&&&&& \\ & 1 &&&&&& \\ && 1 &&&&& \\ &&& 1 &&&& \\ &&&& 1 &&& \\ &&&&& 1 && \\ &&&&&& 1 & \\ &&&&&&& 1 \end{bmatrix} * \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \\ \xi_5 \\ \xi_6 \\ \xi_7 \\ \xi_8 \end{bmatrix} + \delta \qquad (5.3)$$

5.5. Model details

Some elements of the model are discussed below in greater detail.

5.5.1. Fixed λ coefficients

The entries in the matrices Λ_Y and Λ_X represent the structural coefficients linking the latent concepts to their respective observed indicators. It can be argued that that the latent variables are hypothetical (Hayduk 1987), and that they do not have a measurement scale on their own (Hox and Bechger 1998). Therefore, the non-zero entries in the two matrices set the scale on which the values of the underlying concepts are measured.

Since we can measure a latent variable on any scale we wish, it is convenient to fix one λ_Y^{ij} value to 1.0 for each latent variable. In this way, we ensure that a latent concept is measured on the same scale as its observed indicator, and thet a unit change in a latent concept will correspond to a unit change in its observed indicator (Hayduk 1987, Hayduk 1996).

The case of multiple indicators of a concept is addressed separately later in this chapter.

5.5.2. Fixed indicators' measurement errors

The presence of a measurement error δ or ϵ indicates that a latent concept is not measured perfectly by an indicator, and that a portion of an indicator's variances is produced by entities other than the corresponding concept (Hayduk 1987, Hox and Bechger 1998).

Hayduk (1987, 1996) argued that measurement errors (also called "measurement validities") should normally be fixed rather than left free. Fixing a measurement error's variance, firstly, specifies the researcher's familiarity with data collection and recording procedures.

Secondly, by changing the amount of an indicator's error variance, we can change the meaning of the underlying concept (Hayduk 1987) and quantify the similarity between a concept and its indicator (Hayduk 1996). For example, the midterm test score was used as the measurement of the students' knowledge. We can conceptualize students' knowledge as "knowledge gained in a classroom and from self-studying." Than we can assign, say, 10% of an indicator's variance σ_y to the error variance Θ_ϵ, to account for the fact that a midterm test is not a perfect indicator of knowledge gain, and that some other factors (e.g., fatigue or anxiety) might affect the test score. Had we decided to conceptualize student knowledge as "knowledge gained in the classroom only," we would have to assign a higher proportion of an indicator's variance to the error variance, since, in this case, the test score would not reflect the knowledge gain from self-studying. Re-estimating a model with different conceptualizations of a concept may produce a better- (or worse-) fitting model.

Hayduk (1987) made some general suggestions about the relative size of a measurement error. Some physical characteristics, such as gender, will have a low proportion of error variance (Hayduk used 1% of the total indicator variance). When subjects have a tendency to misreport data (as in the case of age or income), or when the question itself has answer categories that contain broad ranges (for example, the test score question had categories with a range of 10 marks), the error variance can be in the vicinity of 5-10% of the indicator's variance. If respondents are asked about matters not encountered routinely in daily life and have to spend time considering the answer, the proportion of error variance will be higher. When a data-entry procedure is itself subject to the error (for example, manual entry versus electronic scanning), the error variance will be even higher.

In the presented SEM model, the indicators' error variances were fixed at the following values (see Table 5.4):

Table 5.4: Proportion of indicators' fixed error variance

Indicator	Proportion of fixed variance, %	Rationale
y_1	10	Some students marked their answer between the categories to report, for example, participating in extra-curricular activities 3 time a week (e.g., marking between *d* and *e*); also, averaging over a period of the semester required approximation.
y_2	20	Some students reported they never speak English even if their background was English-speaking; also, imprecise measurement scale.
y_3	5	"Similarity" of the course was not well-defined.
y_4	free	Multiple indicator of a concept (see Section 4.3.5.3).
y_5	50	Multiple indicator of a concept (see Section 4.3.5.3).
y_6	20	Sample might not be representative of the whole class, since students who miss most of the lectures may have missed the survey as well; averaging over a period of semester that required approximation.
y_7	10	Test score was not reported as a number, but as belonging to a 10-point wide category (to preserve student's anonymity)
y_8	10	Imprecise measurement scale; also, response was based on subjective perception.

Table 5.4 continued

Indicator	Proportion of fixed variance, %	Rationale
y_9	20	Tendency to over- or under-report satisfaction (e.g., report as "black or white" and not in "gray" tones).
y_{10}	20	Tendency to over- or under-report satisfaction (e.g., report as "black or white" and not in "gray" tones).
y_{11}	20	Tendency to over- or under-report satisfaction (e.g., report as "black or white" and not in "gray" tones).
y_{12}	10	As a binary (yes-no) variable, there was no "middle ground" to report an opinion.
x_1	5	Owning a personal computer is not a perfect measure of financial background.
x_2	1	Question about gender is unambiguous, error attributed to data entry.
x_3	5	As a binary (yes-no) variable, there was no "middle ground" to report an opinion.
x_4	5	As a binary (yes-no) variable there was no "middle ground" to report an opinion.
x_5	10	Three answer categories do not provide a precise definition of language background.
x_6	10	Age was recorded not as actual age, but as belonging to a category (to preserve student's anonymity).
x_7	1	Question about graduate status is unambiguous, error attributed to data entry.
x_8	20	Instructor's teaching experience was approximated by years teaching and was coded on an ordinal scale

In some conditions, error variance can be freed and left to be estimated by the model:

- when the quality of measurement and not variable conceptualization is the modeling objective, and
- in dealing with multiple indicators of a concept (discussed later in the chapter) (Hayduk 1987).

Fixing an indicator's measurement error, in fact, indirectly fixes the variances of the exogenous concepts and constrains the variances of the endogenous concepts (the variances of the endogenous concepts arise from the causal actions of coefficients in matrices B and Γ) (Hayduk 1987). This phenomenon can be illustrated in a pictorial (see Figure 5.1) and in equation form. In equation form,

$$Var\,(x_1) = \lambda_{X1}^2\, Var\,(\xi_1) + Var\,(\delta) \tag{5.4}$$

Moreover, since $\lambda_{X1} = 1$, Var (x_1) is a data artifact (a fixed value in matrix *S*), and Var (δ) is fixed by specification, Var (ξ_1) is fixed indirectly as

$$Var\,(\xi_1) = Var\,(x_1) - Var\,(\delta) \tag{5.5}$$

Equation (5.5) highlights that the variance of an indicator will be greater than (or equal to) the variance of a corresponding concept, since some other factors besides the concept may influence the indicator.

In a pictorial form, the phenomenon can be illustrated as follows:

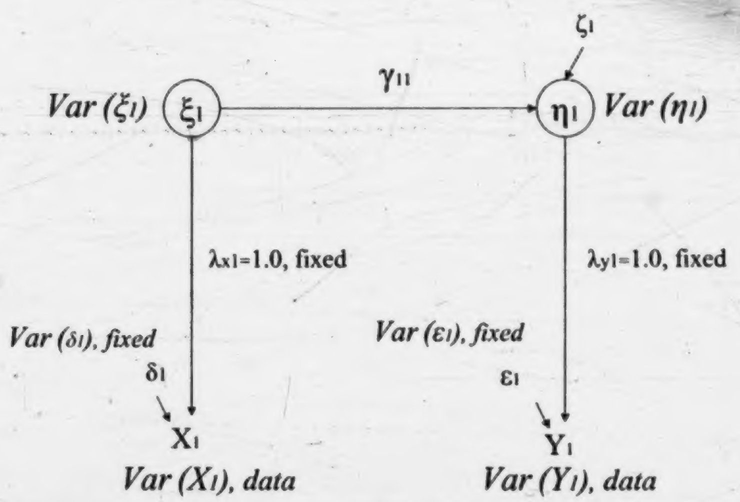


Figure 5.1: SEM model with fixed indicators' error variances

5.5.3. Multiple indicators

To estimate the time students spend on studying outside the classroom, they were asked to report two values: the amount of time per week spent on homework, and how many times a week they were using classroom notes or the textbook.

Both of these values provide an estimate of how much effort/time students devote to self-studying. Both the homework time and reading the notes/textbook appeared to estimate the same latent (unobserved) variable – the time/effort devoted to self-studying. Such a construct can be presented, again, in pictorial and in equation form. The pictorial representation is provided in Figure 5.2 below:

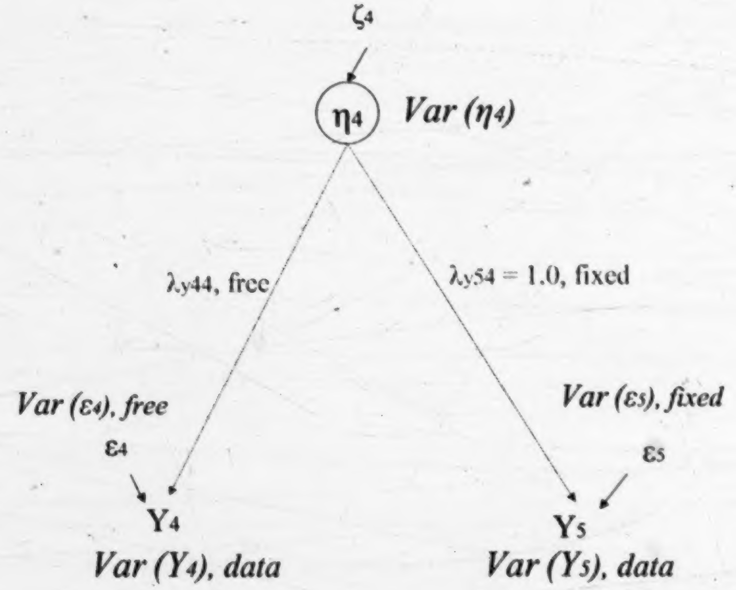

ζ4
η4
Var (η4)
λy44, free
λy54 = 1.0, fixed
Var (ε4), free
ε4
Var (ε5), fixed
ε5
Y4
Var (Y4), data
Y5
Var (Y5), data

Figure 5.2: Latent variable "Time devoted to self-studying" with multiple indicators

Hayduk (1996) suggested fixing the structural coefficient λ to 1.0 for the indicator that is most similar to the unobserved latent variable. In this way, the best indicator gives the latent variable a measurement scale. Furthermore, by fixing the measurement error of the best indicator, the latent variable is conceptualized into whatever the modeler wants it to represent. The rest of the indicators are given free structural coefficients λ and free measurement error variances. Since the latent variable's scale and meaning are already specified by the fixed λ and the measurement error for the best indicator, freeing the coefficients for the rest of the indicators serves as the test of the modeler's conceptualization of it.

In equation form,

$$Var\ (y_5) = Var\ (\eta_4) + Var\ (\epsilon_5) \tag{5.6}$$

$$Var\ (y_4) = (\lambda^{y}_{44})^2\ Var\ (\eta_4) + Var\ (\epsilon_4). \tag{5.7}$$

The value of Var (η_4) in Equation (5.6) is constrained by Var (ϵ_5) and Var (y_5), since those are both fixed values. If Var (ϵ_4) were fixed as well, then λ^{y}_{44} would have been tightly constrained to assume the value that would force the equality of both sides of Equation (5.7).

From the perspective of conceptualizing the meaning of the latent variable "Time devoted to self-studying", neither of the indicators (y_4 – "Reported homework time" and y_5 – "Reported time spent reading text/notes") provided a perfect measure. Each indicator measured a somewhat different concept. Still, it was believed that the two indicators combined to provide a valid estimate of the latent variable "Time devoted to self-studying". It was expected, therefore, that a high proportion of error variance would have to be assigned to one of the indicators, and that the second indicator's model-estimated measurement error would be significant as well.

In this model, the structural coefficient λ^y_{54} connecting the latent variable "Time devoted to self-studying" to the indicator y_5 "Reported time spent reading text/notes" was fixed to provide meaning to the unobserved latent variable. The other structural coefficient, λ^y_{44}, was left free to vary.

5.5.4. Reciprocal effects

In the real world, causality often is a two-way street. For example, students' achievement depends on their motivation (i.e., motivation is a cause of achievement), but motivation, in turn, depends on achievement (i.e., achievement is a cause of motivation) (Richmond et al. 2000). In the physical world, reciprocity is a well-known phenomenon, formulated by Isaak Newton in the form of his famous third law: for every action, there is an equal, and opposite reaction.

A reciprocal relationship between the latent variables "Time devoted to self-studying" and "Student knowledge" was created in the SEM classroom educational model (see Figure 5.3):

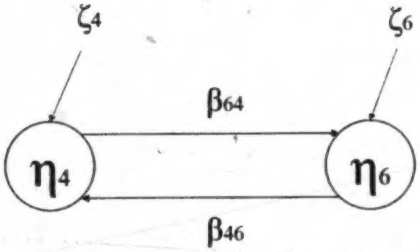


Figure 5.3: Reciprocal effect between latent variables in the SEM model

The reciprocal structure created a feedback effect: a change in the variable η_4 by one unit (coming from the effects of other variables) produced β_{64} change in the variable η_6, which, in turn, produced β_{46} change in η_4. The latest change is again transmitted to the variable η_6 and back, and the cycle goes on, theoretically, forever. If the absolute value of the product of all structural coefficients composing the loop L ($L = \beta_{64} x \beta_{46}$ in our case) is less than 1 (as will be the case if we consider the standardized values of the coefficients), each cycle will provide a change in a variable that is smaller that the change provided by a previous cycle (Hayduk 1987).

In many disciplines dealing with dynamic processes, when L is negative and its absolute value is less than 1, a system will exhibit behavior called "damped oscillations." Also, if $L = -1$, the oscillations become "sustained," and a negative L with an absolute value greater than 1 will produce "amplified" oscillations. A positive L with an absolute value greater than 1 will produce exponential growth, (see, for example, Sterman 2000)). Hayduk (1996) called systems with loops with a L value greater than 1 "explosive."

5.5.5. Loops

An effect may return to a variable through a more complex chain than the two-variable reciprocal system described in the previous section. For example, in this model, the latent variables "Time devoted to self-studying" η_4, "Perceived course workload" η_7, "Satisfaction with the course in general" η_8, and "Attendance" η_5 are involved in the causal loop named the "Workload Effect I" (see Figure 5.4):

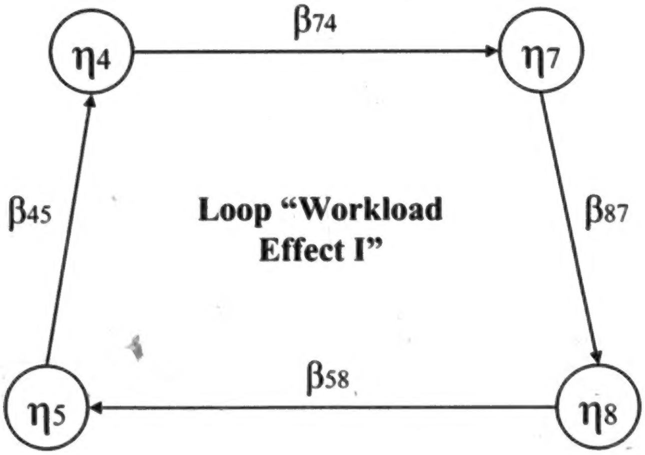


Figure 5.4: Loop "Workload Effect I"

In this loop, an increase in the time devoted to studies will increase the perceived workload, and this increase, in turn, will produce lower satisfaction with the course. Lower satisfaction with the course will increase lecture absenteeism, which, in turn, might increase (for those who want a higher mark), or decrease (for those disenchanted with the course and subject) the time students spend self-studying.

The other causal loops in the SEM model will be discussed in detail in Chapter 6.

5.5.6. Control variables

The model's matrix Γ, reveals that the latent variables ξ_1 ("Financial background") and ξ_2 ("Gender") have no direct causal effects on the other variables. Each of these variables provides no effects for a different reason.

The students' gender was one of the parameters collected during the study. One of the goals was to test the assumption that gender does not influence any other of the model's variables. Specifically, the assumption was that gender does not influence either the knowledge gained or the attitudes toward the course or instructor.

After collecting questionnaires and analyzing responses, it became obvious that owning a personal computer was a poor measure of financial background since most of the respondents (361 out of 384) did own a computer. This fact also turned indicator x_1 into almost a constant.

The mere fact of a high proportion of "yes" responses, and resulting low variance, was not the primary reason for modeling no direct causal effects from the variable "Financial background." For example, the question "Do you believe it is important to have fun while at university" received an even higher proportion of "yes" answers(indicator x_2, 379 "yes" out of 384 answers). However, it was still believed that someone who thought that enjoying time spent in university was important would more actively participate in extra-curricular activities (causal effect from ξ_3 to η_1) than someone else.

Including concepts that provide no direct causal effects (or "control variables") increases a model's discriminatory power (Hayduk et al. 1997). This increase results from the increase in the model's number of degrees of freedom. In this model, adding two exogenous variables with no direct causal effects increased the number of elements in the covariance matrix S by (20 + 19) = 39, but required only (8 + 7) = 15 additional model parameters to be estimated (the variables' own variances and the covariances with other exogenous variables). Thus, the model's total number of degrees of freedom increased by (39 – 15) = 24.

Modelers should seek models with many degrees of freedom (few estimated parameters) (Hayduk 1987). If a model has as many estimated parameters as there are entries in an input matrix S (this equality could be achieved by specifying all the possible paths between all variables), the model would fit the data perfectly, but would be just as complex as the data themselves (Hayduk 1987, Hox and Bechger 1998).

5.6. Model results and assessment of fit

The model "Original" was estimated by using the LISREL 8.30 software. The model's syntax is presented in Appendix XIV.

The results of the chi-square test indicate that the model fits the data adequately (χ^2 = 121.47 with 108 d.f., $P = 0.177$). The software also produces a number of goodness-of-fit indices that, in addition to assessing the fit of the model, assess its simplicity (Hox and Bechger 1998). Some of the widely used indices are the Adjusted Goodness-of Fit (AGFI) and the Root Mean Square Error of Approximation (RMSEA). For this model, AGFI = 0.94, and RMSEA = 0.017. Both indices indicate that the model is acceptable, but an AGFI of 0.95 is normally required to conclude that a model is "good" (Hox and Bechger 1998). For a summary of the discussion on goodness-of-fit indices, the reader can refer to the SEMNET Discussion Network (SEMNET).

The estimates of the structural coefficients β and γ are presented in Table 5.5.

Table 5.5: LISREL estimates of the structural coefficients β and γ, model "Original"

Coeff.	**LISREL Estimate**	**Coeff.**	**LISREL Estimate**	**Coeff.**	**LISREL Estimate**
β_{41}	-0.0621 (Z=-1.50)	$\beta_{10\text{-}8}$	0.3658	γ_{15}	-0.2306
β_{61}	0.3194	β_{59}	-0.1017 (Z=-0.97)	γ_{25}	-0.3840
β_{42}	-0.0079 (Z=-0.15)	β_{89}	0.7274	γ_{45}	0.0595 (Z=0.85)
β_{62}	0.2683	$\beta_{10\text{-}9}$	0.3265	γ_{65}	-0.3791
β_{63}	0.0079 (Z=0.08)	$\beta_{4\text{-}10}$	0.0355 (Z=0.78)	γ_{75}	0.1686
β_{73}	-0.1198	$\beta_{5\text{-}10}$	0.0065 (Z=0.09)	γ_{16}	-0.4147
β_{83}	0.0880 (Z=1.47)	$\beta_{1\text{-}11}$	0.7427	γ_{46}	0.2027
β_{93}	-0.1758	γ_{13}	0.9838	γ_{86}	0.1055 (Z=1.04)
β_{64}	1.4538	γ_{73}	0.6181 (Z=1.78)	γ_{96}	0.4339
β_{74}	0.3274	$\gamma_{11\text{-}3}$	0.2449 (Z=1.36)	γ_{17}	-0.0458 (Z=-0.22)
β_{45}	-0.1925	γ_{44}	0.4176	γ_{37}	0.3113
β_{65}	0.0711 (Z=0.33)	γ_{54}	-0.7940	γ_{47}	0.3739
β_{46}	-0.1860 (Z=-1.77)	γ_{98}	0.2438	γ_{57}	-0.2463 (Z=-1.79)
β_{86}	0.2431	γ_{58}	-0.0881 (Z=-1.87)	γ_{67}	0.0851 (Z=0.19)
β_{96}	0.0894 (Z=1.49)	$\gamma_{11\text{-}4}$	-0.0011 (Z=-0.01)	γ_{87}	0.1140 (Z=0.61)
β_{87}	-0.0904 (Z=-1.42)	$\gamma_{10\text{-}8}$	-0.0600 (Z=-1.05)	γ_{97}	0.4659 (Z=1.88)
β_{97}	-0.2731	γ_{88}	-0.0156 (Z=-0.28)	$\gamma_{10\text{-}7}$	0.2255 (Z=1.35)
β_{58}	0.0011 (Z=0.009)			γ_{48}	0.0885

For the statistically non-significant coefficients, the Z values are given in parentheses.

5.6.1. Looking beyond fit indices

For several reasons, when evaluating a model's quality, one must look beyond the fit indices. Firstly, as with any statistical test, we cannot claim that we found the *right* model. It is possible to specify a different, but equally acceptable (so-called "equivalent") models for the same data set (Hayduk 1996). The proper conclusion about a fitting model is that a model and set of coefficient estimates that are consistent with the observed covariances has been located (Hayduk 1987). One can find parallels in the statistical hypothesis testing, where "non-rejection" of a hypothesis does not imply "acceptance" (Montgomery and Runger 1999).

Secondly, even if the indices indicate an acceptable fit, the model might still be inappropriate because of the wrong signs of coefficients, a miniscule R^2, significant or non-normally distributed residuals, or some other improprieties (Hayduk 1996). LISREL output provides diagnostics data that can be used to assess the appropriateness of specific model elements.

5.6.2. Squared Multiple Correlations

The Squared Multiple Correlations reported by LISREL for the latent variables and indicators are analogous to the "coefficient of determination" r^2 in regression analysis:

$$r^2 = (\mathit{Regression\ Sum\ of\ Squares}) / (\mathit{Total\ Sum\ of\ Squares}). \qquad (5.8)$$

The coefficient of determination shows the proportion of variation in the dependent variable explained by the regression model (Montgomery and Runger 1999, Levine et al. 1998). In SEM, the squared multiple correlations bear the same meaning – the proportion of variation in a variable explained by the model:

$$R^2 = (explained\ variance) / (total\ variance) = 1 - (error\ variance) / (total\ variance) \qquad (5.9)$$

Note that while the term "variation" or "variability" is used in defining r^2 in regression analysis (Montgomery and Runger 1999, Levine et al. 1998), the term "variance" is used in defining R^2 in structural equation modeling (Hayduk 1996). Both "variation" and "variance" account for the same phenomenon – the amount of dispersion in the data. The difference between the two is that in computing the variance, the sum of squared differences around the mean is divided by the sample size (n) minus 1:

$$Variation \equiv Total\ Sum\ of\ Squares = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (5.10)$$

$$Variance = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}. \qquad (5.11)$$

Explaining the variation in endogenous variables is one of the goals of modeling. Table 5.6 presents the squared multiple correlations for these variables.

Table 5.6: Squared multiple correlations for latent endogenous variables

SEM Variable	**Squared Multiple Correlation R^2**
η_1	0.199
η_2	0.087
η_3	0.016
η_4	0.269
η_5	0.170
η_6	–0.065
η_7	0.114
η_8	0.777
η_9	0.210
η_{10}	0.534
η_{11}	0.006

The squared multiple correlation for variable η_6 is the negative 6.5%! Was this a result of a software bug? Hayduk (1996) demonstrated that, in fact, the traditionally used formula for R^2 (Equation (5.9)) is misleading when used for computing the squared multiple correlation for variables affected by loops (or reciprocal relations, which are loops as well).

The value of –0.065 comes from the Equation (5.9):

$$R^2 = 1.0 - (\text{proportion of error variance}) = 1.0 - \text{Var}(\zeta_6) / \text{Var}(\eta_6) = 1.0 - 1.340 / 1.257 = -0.065.$$

When an error variable (ζ_6 in this case) contributes its error variance to a latent variable affected by a loop, the total contribution of the error variable differs from the coefficient 1.0 implied by the model. The error variable itself is also affected by a loop (Hayduk 1996). To account for a loop's effects on an error's variable contribution to the latent variable, Hayduk (1996) suggested computing a *Loop Adjusted R^2*, or *LAR^2*:

$$LAR^2 = 1.0 – (\text{enhanced error variance}) / (\text{total variance}), \qquad (5.12)$$

where the enhanced error variance is the enhanced contribution of the error variable to the dependent variable.

Using the procedure described in (Hayduk 1996, p. 120), one can compute LAR^2 for η_6 by using model output as follows:

$$L_s = 1 – 1/(1 + TE_{\eta\eta \text{diag}}) = 1 – 1/(1 – 0.2056) = -0.259$$

$$LAR^2 = 1.0 – ((1/(1-L_s))^2 \text{ x } Var (\zeta_6)) / Var (\eta_6) = 1 – ((1/(1+0.259))^2 \text{ x } 1.340) / 1.257 = 0.327.$$

Therefore, the true amount of the explained variance in variable η_6 is 32.7%. Computing LAR^2 instead of R^2 might be necessary for all the variables involved in loops. Hopefully, in the future, LAR^2 statistics will be routinely included in model output produced by the LISREL software.

For the indicator variables with a fixed amount of error variance, the squared multiple correlation reports merely the proportion of variance allocated to the variable by a modeler. For example, for the indicator x_6 ("Reported Age"), 10% of the total indicator's variance was specified as the error variance. The LISREL output reported $R^2 = 0.90$ for that indicator.

5.6.3. Multiple indicators

The LISREL estimates for the variable η_4 with indicators y_4 and y_5 are presented in Figure 5.5:

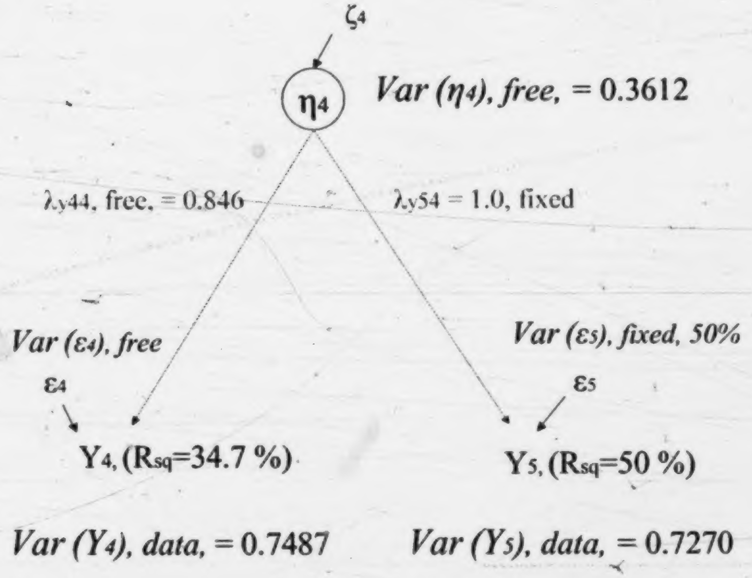

ζ4
η4
Var (η4), free, = 0.3612
λy44, free, = 0.846
λy54 = 1.0, fixed
Var (ε4), free
ε4
Var (ε5), fixed, 50%
ε5
Y4, (Rsq=34.7 %)
Y5, (Rsq=50 %)
Var (Y4), data, = 0.7487
Var (Y5), data, = 0.7270

Figure 5.5: LISREL estimates for variable η_4 with multiple indicators

The imprecise nature of the pair of indicators as a measure of the latent variable "Time devoted to self-studying" required assignment of a 50% of the total variance of indicator y_5 to the error variance, and the LISREL software estimated that the error variance of indicator y_4 amounted to 65.3% of the indicator's total variance (The amount of explained variance in indicator y_4 can be computed as $\mathrm{Var}(\eta_4)*(\lambda^{y}_{44})^2$, or $R_4^2 = 0.3612*0.846^2 = 0.2585$, or 34.7% of $\mathrm{Var}(y_4)$). These estimates confirmed the suspicion that each of the indicators was measuring a variable somewhat different from the "Time devoted to self-studying."

The value of λ^{y}_{44} of 0.846 appears because of the approximate equality of the variances of both indicators. The correlation 0.41 between indicators y_4 and y_5 might suggest that indicators do not function well as measures of the same concept. The analysis of the standardized residuals and modification indices associated with both indicators suggested, however, that the only problematic point was the association of the indicators y_4 and y_5 with the indicator y_8 ("Reported perceived workload"). It was decided, therefore, to keep the multiple indicators y_4 and y_5 in the model.

5.6.4. Analysis of residuals

LISREL computes residuals as the differences between the individual elements of the data matrix S and the model-implied matrix Σ. While a matrix of the absolute values of residuals is available, LISREL also provides a matrix of the standardized residuals – the estimates of the number of standard deviations the observed residuals are away from zero (the standard assumption is that residuals are normally distributed with a mean of zero). Standardized residuals greater than +2 or smaller than –2 deserve special attention (Hayduk 1987).

Residuals can also be analyzed by looking at the patterns of their behavior to test the assumption of normality. LISREL provides a so-called Q-plot (also known as the "normal probability plot" (Levine et al. 1998)). If the residuals were distributed normally, they would plot along the 45° line on a Q-plot. Deviations from a straight line indicate non-normality, and a straight-line pattern with the slope of the line steeper (or gentler) than 45° indicates that the residuals are distributed less variably (more variably) than would be expected based on the asymptotic variances used to standardize the residuals (Hayduk 1987).

An analysis of the model's standardized residuals indicated a string of statistically significant residuals involving the indicator of the instructor's experience (x_8, "Years teaching"). Substantial residuals (with absolute values between 2.39 and 3.66) were for the covariances between x_8 ("Years Teaching") and y_8 ("Reported perceived workload"), x_8 ("Years Teaching") and y_9 ("Reported satisfaction with the course"), x_8 ("Years Teaching") and y_{10} ("Reported satisfaction with the instructor"), and x_8 ("Years Teaching") and y_{11} ("Reported attitude toward the subject"). Statistically significant residuals indicate that the model failed to account for the degree of coordination between the respective indicators.

An analysis of matrix Γ might provide some insight into the residuals' origin. The structural coefficient γ_{78} between the latent variables "Instructor's teaching experience" (ξ_8) and "Perceived course workload" (η_7) was fixed at zero, postulating no direct causal relationship. The structural coefficients γ_{88} between the latent variables "Instructor's teaching experience" and "Satisfaction with course" (η_8), and γ_{10_8} between "Instructor's teaching experience" and "Attitude toward the course subject" (η_{10}) were statistically non-significant, thus failing to account directly for the coordination between the respective indicators.

On a Q-plot, the residuals fell on an approximately straight line at 45°, indicating that the residuals were indeed distributed normally with variability in accordance with the expected asymptotic variance.

5.6.5. Reciprocal effects and correlation among the estimates

A high degree of correlation between two model coefficients indicates that the estimate of one coefficient is strongly related to the estimate of the other, or, in other words, that colinearity exists between two estimates. Colinearity might lead to the failure to identify unique values for the parameters (i.e., an increase in the value of one coefficient will be offset by a decrease in that of another, if the two are strongly negatively correlated) (Hayduk 1987).

The correlation between the reciprocal effects β_{64} and β_{46} of –0.9015 indicates that the two parameters might be collinear: more than one set of parameter values might imply the same matrix Σ. Colinearity between reciprocal effects is a typical problem. The solution to the identification (colinearity) problem is to place additional constraints on the effects. This solution can be achieved by, for example, introducing an exogenous variable that strongly influences one reciprocally related variable, but not the other (Hayduk 1987).

The adjustments made to the model in an attempt to resolve the colinearity issue, along with other changes to the model, are described in the next section.

5.7. Model Improvement

LISREL output provides diagnostics data that can be used to improve model fit. For the fixed coefficients, the software computes the partial derivatives and modification indices. Partial derivatives indicate the slope of the fit function at the current value of the parameter. If the slope is not equal to zero, the function has not reached the minimum. Since partial derivatives are computed in real metrics (e.g., points of test score), determining what size of partial derivative is large enough to substantially improve the model fit (by decreasing chi-square), is a difficult task (Hayduk 1987).

Therefore, evaluating so-called modification indices for each fixed parameter could be more informative. Modification indices are based on partial derivatives (Hayduk 1987), and the statistical test used is called the "Lagrange multiplier test" (Hox and Bechger 1998). The modification index reports the minimum expected decrease in the chi-square statistic (thus reducing the difference between matrices S and Σ), if a fixed coefficient is freed. Freeing one coefficient reduces a model's degrees of freedom by one. The difference between the "old model's" chi-square and the "freed-model's" chi-square is also a chi-square variable with one degree of freedom. The value of the "difference" chi-square statistic is statistically significant at 0.05 level if it exceeds 1.96^2, or 3.84. Therefore, freeing a coefficient with a modification index greater than 4.0 indicates that the model would improve statistically significantly.

However, a caveat must be made about relying on modification indices to free fixed coefficients (especially those fixed at zero value), or to delete statistically non-significant parameters. Model adjustments based solely on the analysis of output amounts to fitting a model to the data, even though the goal of modeling is to test one's theory (or test someone else's theory) (Hayduk 1987). Fitting a model to the data also presents the danger of capitalizing on the random properties of the sample (Hox and Bechger 1998, Hayduk 1996). Model adjustments should have theoretical justifications (Hayduk 1987). Since justification is often introduced retrospectively, researchers might become "very creative in justifying modifications" (Hox and Bechger 1998, p. 9).

5.7.1. Adjustments to the SEM model: non-significant coefficients

The analysis of the B and Γ matrices revealed a number of statistically non-significant structural coefficients β and γ. It was decided to remove some of them and to leave the others. The decision-making process, on which the removing/leaving determination was based, proceeded along the following considerations. If some effects were truly believed to belong to a model, despite their statistical non-significance, they were left in the model. Non-significance could have been caused by random sampling (such as in case of $\beta_{4\text{-}10}$), or an effect might become significant once changes are introduced (such as in case of β_{58}). On the other hand, non-significance could have been a sign that an effect indeed did not belong in the model (such as in case of β_{63}). Such effects were removed from the model. Table 5.7 lists the reasons for removing effects from the model.

Table 5.7: Structural effects removed from the SEM model

Effect	**Reason for removing**
β_{42}	Not speaking English in everyday life does not affect the time spent on assignments and reading text/notes because non-native speakers have a good command of reading English.
β_{63}	Taking a similar course previously will not give a student an advantage on an exam, since instructors design tests to test knowledge gained during the course only, and not knowledge brought from the outside.
$\beta_{5\text{-}10}$	Lecture attendance is governed by the attitude toward the course and instructor. E.g., even if a student likes the subject, but does not like the course and the instructor, attendance will not improve.
$\gamma_{11\text{-}4}$	Those who believe in importance of doing well in university still find time for fun. Those who do not believe it is important to do well in university will definitely have time for fun.
γ_{45}	By the time students reach their 3rd or 4th year in university, non-native English speakers develop a good grasp of English listening comprehension.
γ_{17}	Graduate status by itself does not affect participation in extra activities when we control for age and background.

5.7.2. Adjustments to the SEM model: modification indices

The analysis of the modification indices revealed that a significant improvement in model fit could be achieved by freeing some of the fixed coefficients (see Table 5.8). Since a modeler should avoid modifying a model just to improve fit (Hayduk 1987), it was decided to analyze whether the changes suggested by modification indices were theoretically sound.

Effect β_{18} indicated that the model did not explain the coordination between the latent variables η_8 "Satisfaction with the course in general" and η_1 "Extra-curricular activities." A direct causal effect from satisfaction to participation in extra-curricula activities appeared unfeasible, so this effect was not added to the model. Effect β_{19} was not added to the model for the same reason.

Table 5.8: Statistically significant modification indices

Coefficient	Required Effect	Modification Index
β_{18}	Fr η_8 to η_1	5.73
β_{78}	Fr η_8 to η_7	6.44
β_{19}	Fr η_9 to η_1	4.24
β_{79}	Fr η_9 to η_7	8.44
γ_{78}	Fr ξ_8 to η_7	12.90
γ_{82}	Fr ξ_2 to η_8	4.10
γ_{92}	Fr ξ_2 to η_9	6.71
ψ_{19}	Cov between ζ_1 and ζ_9	4.44
$\Theta\epsilon_{84}$	Cov between ϵ_4 and ϵ_8	4.62
$\Theta\epsilon_{85}$	Cov between ϵ_5 and ϵ_8	10.43

Effect β_{78} suggested a causal effect from "Satisfaction with the course in general" to "Perceived course workload". The model contained an effect running in the opposite direction: from perceived workload to satisfaction with the course. It was still believed that perceived workload depended primarily on the amount of time devoted to studying and not on the degree of satisfaction with the course. Effect β_{79} was not added to the model for the same reason.

Effect γ_{78} suggested the strong influence of "Instructor's teaching experience" on "Perceived course workload". This relation appeared feasible – an inexperienced instructor may assign an excessive or insufficient amount of homework, or make tests too easy or too difficult. Nonetheless, it was believed that such a relation has to manifest itself through the chain "instructor's experience – time devoted to self-studying – perceived workload."

Effects γ_{82} and γ_{92} suggested that a student's gender influenced his or her satisfaction with the course and the instructor. If freed, these coefficients would have negative values, meaning that females would rate the course and the instructor lower than males.

The high modification index for effect ψ_{19} was another manifestation of the model's inability to account for the coordination between the latent variables "Extra-curricular activities" and "Satisfaction with the course in general". The modification indices for effects $\Theta\epsilon_{84}$ and $\Theta\epsilon_{85}$ suggested correlations between the error variances of indicators "Years teaching" (x_8) and "Reported importance of doing well in university" (x_4), and "Years teaching" (x_8) and "Reported language background" (x_5). These modification indices likely arose from the general problems with the indicator of instructor experience (a number of statistically significant residuals were associated with it).

No significant modification indices were found for the concept ξ_1 "Financial background." This result indicated that it indeed acted as a control variable.

All the modification indices were believed to have arisen out of random sampling fluctuations in the data and, therefore, did not warrant adjustments to the model. None of the direct effects suggested by modification indices were included in the adjusted model.

5.8. Adjusted model: results and discussion

The only changes introduced into the model were removal of the six structural coefficients specified in Table 5.7. This modification added six degrees of freedom to the model. The adjusted model, named "Final", is presented in Appendix XIII. The model produced a chi-square statistic of 122.28 (114 d.f., P = 0.281), which indicated an improved fit. The goodness-of fit indices were AGFI = 0.94, and RMSEA = 0.012. Table 5.9 presents the comparison of the values of the structural coefficients β and γ between the original and the adjusted models.

None of the effects changed signs. The largest changes in the structural coefficients were increases in β_{46} (27.4%), β_{64} (20.7%), γ_{47} (18.9%), and β_{86} (11.4%). The analysis of the residuals and the modification indices revealed no substantial changes from the original model.

The correlation of –0.9015 between the reciprocal effects β_{64} and β_{46} in the original model suggested that the estimates might have been collinear. The same effects exhibited a correlation of –0.8579, which represented a modest improvement. The removal of several direct causal effects leading to the variables η_4 and η_6 might have produced this effect.

It was decided that the adjusted model satisfied the objective postulated in Section 4.3.1. – to illustrate how a classroom educational system can be modeled by using the Structural Equation Modeling approach and LISREL software, with data provided by a classroom educational performance framework. No further model modifications were, therefore, attempted.

Table 5.9: Structural coefficients β and γ values for the original and adjusted models

Coeff	Model 'Original'	Model 'Final'	Coeff	Model 'Original'	Model 'Final'	Coeff	Model 'Original'	Model 'Final'
β_{41}	-0.0621 (Z=-1.50)	-0.0579 (Z=-1.37)	β_{58}	0.0011 (Z=0.009)	-0.0027 (Z=-0.02)	γ_{65}	-0.3791	-0.3840
β_{61}	0.3194	0.3558	β_{10-8}	0.3658	0.3577	γ_{75}	0.1686	0.1712
β_{42}	-0.0079 (Z=-0.15)	Removed	β_{59}	-0.1017 (Z=-0.97)	-0.0951 (Z=-0.93)	γ_{16}	-0.4147	-0.4255
β_{62}	0.2683	0.2813	β_{89}	0.7274	0.7279	γ_{46}	0.2027	0.2067
β_{63}	0.0079 (Z=0.08)	Removed	β_{10-9}	0.3265	0.3332	γ_{86}	0.1055 (Z=1.04)	0.1039 (Z=1.03)
β_{73}	-0.1197	-0.1197	β_{4-10}	0.0355 (Z=0.78)	0.0437 (Z=0.89)	γ_{96}	0.4339	0.4344
β_{83}	0.0880 (Z=1.47)	0.0882 (Z=1.47)	β_{5-10}	0.0065 Z=0.09	Removed	γ_{17}	-0.0458 (Z=-0.22)	Removed
β_{93}	-0.1758	-0.1763	β_{1-11}	0.7427	0.7413	γ_{37}	0.3113	0.3114
β_{64}	1.4538	1.7578	γ_{13}	0.9838	0.9810	γ_{47}	0.3739	0.4430
β_{74}	0.3274	0.3346	γ_{73}	0.6181 (Z=1.78)	0.6210	γ_{57}	-0.2463 (Z=-1.79)	-0.2428
β_{45}	-0.1925	-0.2008	γ_{11-3}	0.2449 (Z=1.36)	0.2445 (Z=1.39)	γ_{67}	0.0851 (Z=0.19)	-0.0765 (Z=-0.18)
β_{65}	0.0711 (Z=0.33)	0.1382 (Z=0.64)	γ_{44}	0.4176	0.4393	γ_{87}	0.1140 (Z=0.61)	0.1161 (Z=0.63)
β_{46}	-0.1860 (Z=-1.77)	-0.2374	γ_{54}	-0.7940	-0.7930	γ_{97}	0.4659 (Z=1.88)	0.4694
β_{86}	0.2431	0.2417	γ_{11-4}	-0.0011 (Z=-0.01)	Removed	γ_{10-7}	0.2255 (Z=1.35)	0.2273 (Z=1.36)
β_{96}	0.0894 (Z=1.49)	0.0831 (Z=1.35)	γ_{15}	-0.2306	-0.2387	γ_{48}	0.0885	0.0897
β_{87}	-0.0904 (Z=-1.42)	-0.0902 (Z=-1.41)	γ_{25}	-0.3840	-0.3840	γ_{58}	-0.0881 (Z=-1.87)	-0.0886
β_{97}	-0.2731	-0.2759	γ_{45}	0.0595 (Z=0.85)	Removed	γ_{88}	-0.0156 (Z=-0.28)	-0.0159 (Z=-0.28)
			γ_{10-8}	-0.0600 (Z=-1.05)	-0.0600 (Z=-1.05)	γ_{98}	0.2438	0.2444

5.9. Summary

Chapter 5 presented a cause-and-effect analysis of the classroom system. The system was modeled by using the SEM approach and estimated by using LISREL software. The SEM approach provided modeling opportunities not available in ordinary regression analysis: the separation of the observed indicators from the unobserved theoretical concepts, the specification of the measuring reliability, the inclusion of multiple indicators for a single concept, and the inclusion of loops in the model.

Model output analysis confirmed the presence of feedback structures in a classroom system. While SEM provided insight into the causal world connecting the system's variables, the analysis of the changes to the model variables will be carried out using the Systems Dynamics approach, which is the subject of the next chapter.

Chapter 6: Design and Testing of Policies by Using System Dynamics

6.1. Introduction

The SEM model discussed in the previous chapter allowed us to gain insight into the causal world of a classroom educational system. The model also suggested that some of the variables are involved in feedback loops. Since SEM models do not consider how the behaviour of variables changes over time, a SD model will be created in this chapter to demonstrate the dynamics of the changes in such variables as attitude, satisfaction, and perceived workload. Policies aimed at improving classroom performance will be designed and tested as well.

6.2. System Dynamics modeling steps

No sure recipe is available for building a SD model, but a structured approach aimed at creating a useful model should include the following steps (Richmond et al. 2000, Sterman 2000, Kelton et al. 1998, Andersen and Richardson 1980):

1. Problem definition
    a. Selection of key variables
    b. Selection of time horizon
    c. Selection of model boundary
2. Development of Dynamic Hypothesis
    a. Development of feedback structure
    b. Development of causal loop diagram
    c. Development of stock-and-flow map

3. Model Formulation
    a. Translation of the model into equation form
    b. Specification of model parameters and start values
4. Model Testing
5. Design and Testing of Policies

The steps are presented as a sequence, but iteration can occur from any one step to any other, and the cycle can be repeated numerous times to refine a model created in previous iterations (Richmond et al. 2000, Sterman 2000).

A detailed description of each modeling step is provided in Appendix XV.

6.3. SD Model

The dynamic model of a classroom educational system was built on the insight gained into the system through the application of SEM. The model will be created by using the approach outlined in Section 6.2.

6.3.1. Problem definition

The purpose of this modeling effort was to improve understanding of the dynamic nature of the classroom system and to improve the students' knowledge gain and attitudes toward the subject by introducing of several changes to the course structure.

The Faculty of Engineering at the University of Alberta, where this study was conducted, had, at the time, a policy requiring all engineering students take one engineering management course during their undergraduate program. Therefore, an instructor teaching an engineering management course had only one semester to transfer his or her knowledge to a student in the class and to influence the student's attitude toward the subject of engineering management. A typical semester was 13 weeks long, and, therefore, the model's time horizon was set at 13 weeks as well.

In a SD model, dynamic behaviour should arise from within the system; i.e., such behaviour should be endogenous to the system. In drawing the model's boundary, the model's variables had to be separated into endogenous and exogenous variables. In Chapter 5, a SEM model of classroom performance was created with nineteen variables divided into eight exogenous and eleven endogenous variables. The SEM model provided a reasonable fit for the model's postulated structure; therefore, it was decided to create a SD model based on the same set of variables.

Variables treated as exogenous (such as "Age" or "Language background") were excluded from the SD model. The addition of exogenous variables will not change the model's dynamics. Exogenous variables are not involved in feedback loops, and their effect can be modeled by adding a constant to a variable receiving an effect from an exogenous variable.

6.3.2. Development of a dynamic hypothesis

The SEM model contained a reciprocal relationship between the variables "Time devoted to self-studying" and "Student knowledge," and a number of loops involving various endogenous variables. Each loop represented a feedback structure producing a dynamic behaviour in loop variables. Loops from the SEM model were used to postulate a dynamic hypotheses about the behaviour of educational system variables in the SD model. Each SEM loop was translated into a causal diagram, and all loops were combined into a causal SD model (in creating causal SD loops, SEM variables' Greek names were retained for simplicity). These steps are described below.

6.3.2.1. Balancing loop "Studying"

The variables "Time devoted to self-studying" and "Student knowledge" were involved in a reciprocal relationship in the SEM model. Such a relationship provided a basis for a dynamic hypothesis: as time spent on studying increases, a student's knowledge increases as well. The score on the midterm test measured the level of the student's knowledge. If the midterm test mark was above the expected mark, it was assumed that the student would reduce time spent on studying. The SEM reciprocal construct "Time devoted to self-studying" – "Student knowledge" was translated into a SD balancing loop that was named "Studying" (Figure 6.1) (a balancing loop is a loop in which a change in one variable triggers a sequence of actions that counteract the initial change in the opposite direction).

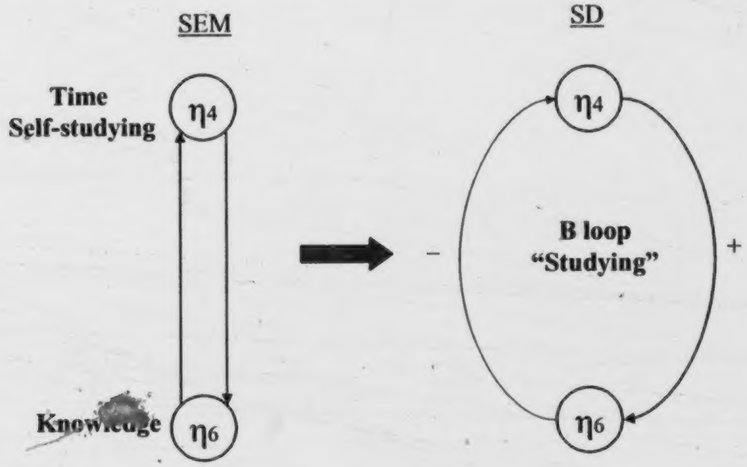


Figure 6.1: Balancing loop "Studying"

6.3.2.2. Balancing loops "Workload effect I" and "Workload effect II"

The SEM variables "Time devoted to self-studying" η_4, "Perceived course workload" η_7, "Satisfaction with the course" η_8, and "Attendance" η_5 created a feedback loop structure (see Section 5.4.5.). This structure was transcribed into the SD causal loop "Workload effect I" (see Figure 6.2).

In this feedback loop, as the time devoted to studying increases, a student perceives a higher workload level. A higher course workload causes a decrease in satisfaction with the course, and this decrease, in turn, results in a higher rate of absenteeism from the lectures. This higher rate also causes a reduction in the time devoted to studying.

A similar loop was providing a balancing (counteracting) effect on "Time devoted to self-studying" through the "Satisfaction with instructor" η_9. This loop was named "Workload effect II" (see Figure 6.2).

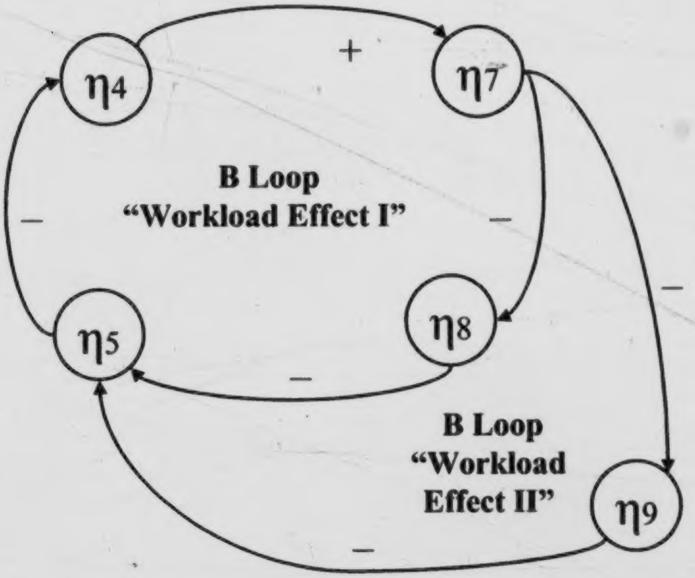


Figure 6.2: Balancing loops "Workload effect I" and "Workload effect II"

6.3.2.3. Reinforcing loops "Attitude effect I" and "Attitude effect II"

While an increase in the time spent on studying increases workload perception, this increase also increases a student's knowledge and, consequently, his or her midterm test mark. A higher test mark improves satisfaction with both the course and instructor, and higher levels of satisfaction improve a student's attitude toward the subject. Students who like a subject are likely to spend more time than other students exploring additional aspects of the course both in and outside the classroom. The described casual structures represent a reinforcing feedback loop: the initial change in one variable is advanced in the same direction by action of the loop. The described loops were named "Attitude effect I" and "Attitude effect II" (Figure 6.3).

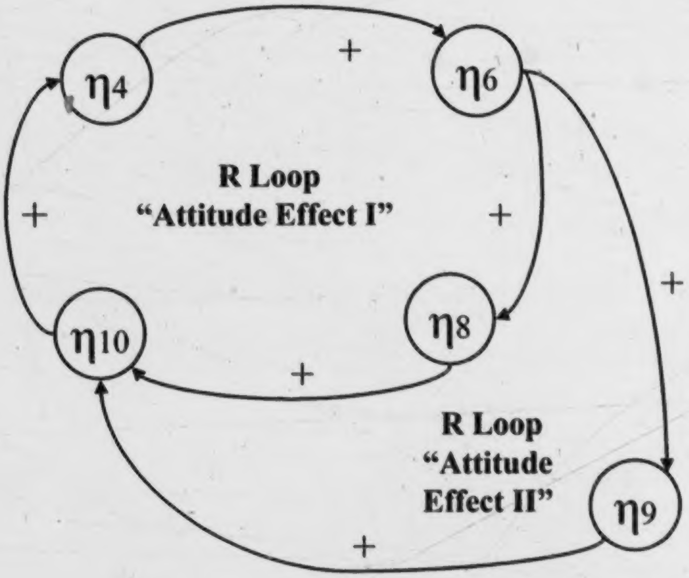
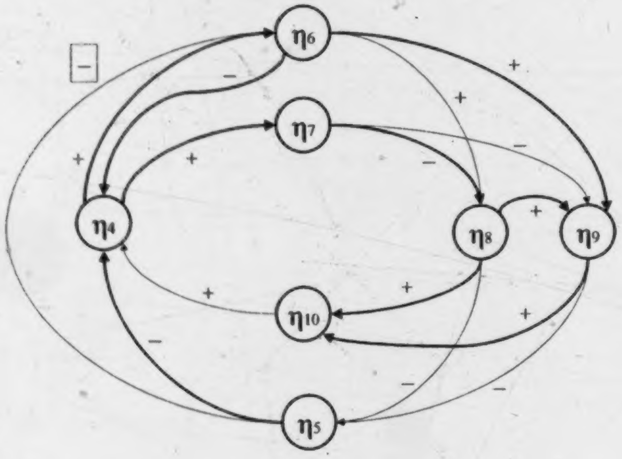


Figure 6.3: Reinforcing loops "Attitude effect I" and "Attitude effect II"

6.3.2.4. Complete causal model

The previous sections illustrated the effects of individual loops on individual variables. In this model, several variables are involved in more than one causal loop. The complete causal model including all variables involved in feedback loops is presented in Figure 6.4.



Variable	Label
Time devoted to self-studying	η_4
Attendance	η_5
Student knowledge	η_6
Perceived course workload	η_7
Satisfaction with the course in general	η_8
Satisfaction with the instructor	η_9
Attitude toward the course subject	η_{10}

Figure 6.4: Complete causal SD model

In Figure 6.4, the thin lines represent structural coefficients that were statistically insignificant in SEM model. The negative sign for the effect from η_5 to η_6 was boxed to highlight that coefficient β_{65} should be positive (and statistically non-significant) in the SEM model. It was believed that the positive sign was a result of sampling fluctuation, and that the true sign of the structural coefficient, representing the effect of lecture absenteeism on knowledge, should be negative: the more lectures that are missed, the less knowledge is gained.

While we can analyze the effect of change in one variable through the action of one causal loop, the combined effect of several loops cannot be easily predicted. A loop with strong causal relationships may dominate the dynamics of the system. As well, loop dominance can shift over time, depending on the state of the model's stocks (Sterman 2000). The model also cannot be easily solved by using analytical methods. Therefore, one must turn to simulation to produce system output and to observe the behaviour of the model's variables.

6.3.3. Model construction

To illustrate the model-construction process, the loop "Studying" will serve as the first building block for the complete model. Model creation will be carried out by using STELLA software – a computer program used for building and running SD models.

6.3.3.1. Stock-and-flow map based on the loop "Studying"

The first step of model construction requires translating a causal loop model into a stock-and flow model, which can be simulated by STELLA software. In the STELLA stock-and-flow model, a rectangle represents a stock (see Figure 6.5), and an arrow originating in a cloud and ending in a stock represents a flow (rate). A cloud indicates that the origin of the flow is outside the model's boundary. A connector (thin arrow) represents an information link (i.e., information input or output). A circle represents an auxiliary variable, which can be a constant, or a converter – a variable specifying a relationship, in equation or graph form (Richmond et al. 2000).

In developing a stock-and-flow model, one decide which variables represent stocks, and which variables represent flow. A helpful approach is to ask oneself the question "What is flowing through the system, and where is the flow accumulated?" Another approach is to "freeze" the system – the accumulators (stocks) should remain in the "frozen" system, and the rates should disappear.

The two stocks identified for the system based on the loop "Studying" were "Real homework time" and "Relative knowledge." The "Real homework time" stock is the actual amount of time per week a student spends on completing homework assignments. The normal homework time spent on assignments per week depends on their rate and difficulty.

If a student spends less time studying than required (the difference is computed by a converter "HW time difference"), the student's rate of knowledge gain will be lower than the rate required to achieve 100 percent on a test. For each hour of homework below the norm, a student will gain β_{46} fewer units of knowledge (stored in the stock "Relative knowledge"). If the student's amount of knowledge is lower relative to the required level, the student's test mark (calculated in the "Test score" converter) will drop. If the test score drops below the "Expected score," the student will compensate by spending more time on homework. Each point of the test score below the expected score will force the student to increase his or her amount of homework time per week by β_{64} hours.

A stock-and-flow model based on the loop "Studying" represented a simplification of the real system. It was assumed that knowledge gain comes only from self-study, that knowledge is not lost through forgetting, and that a student's knowledge is tested continuously.

The STELLA stock-and-flow model based on the loop "Studying" is presented in Figure 6.5.

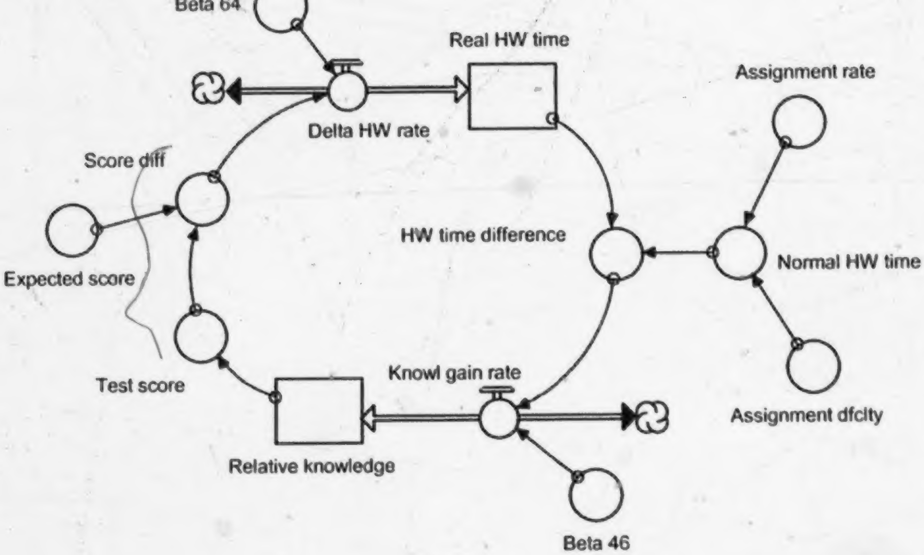

Beta 64
Real HW time
Assignment rate
Delta HW rate
Score diff
HW time difference
Normal HW time
Expected score
Test score
Knowl gain rate
Assignment dfclty
Relative knowledge
Beta 46

Figure 6.5: STELLA stock-and-flow model based on the loop "Studying"

6.3.3.2. Model equations and starting values

A stock-and-flow model has to be translated into a set of equations, and initial values have to be assigned to the stocks and constants.

In this model, it was assumed that the assignment rate was 1 per week, and that the required effort was 3 hours of homework per assignment. The expected test score was set at 80, representing an expectation for a "B" course grade. The "Real HW time" was given an initial value of 1 hour of homework per week, and the initial value of "Relative knowledge" was set at zero. The values of β_{46}. and β_{64} represented the LISREL estimates of the structural coefficients. The value of the test score was bounded by the [0, 100] interval.

STELLA automatically creates a set of equations from the specified stock-and-flow diagram. The equations for the model based on the loop "Studying" are presented below.

```
Real_HW_time(t) = Real_HW_time(t - dt) + (Delta_HW_rate) * dt
INIT Real_HW_time = 1

INFLOWS:
Delta_HW_rate = Beta_64*Score_diff
Relative_knowledge(t) = Relative_knowledge(t - dt) + (Knowl_gain_rate) * dt
INIT Relative_knowledge = 0

INFLOWS:
Knowl_gain_rate = Beta_46*HW_time_difference
Assignment_dfclty = 3
Assignment_rate = 1
Beta_46 = 1.7578
Beta_64 = -0.2374
Expected_score = 80
HW_time_difference = Real_HW_time-Normal_HW_time
Normal_HW_time = Assignment_dfclty*Assignment_rate
Score_diff = Test_score-Expected_score
Test_score = max(min(Relative_knowledge+100,100),0)
```

6.3.3.3. Model behaviour

The behaviour of the model's variables was simulated over a period of 13 weeks. The behaviour of the variables "Test score" and "Real homework rate" is illustrated in Figure 6.6.

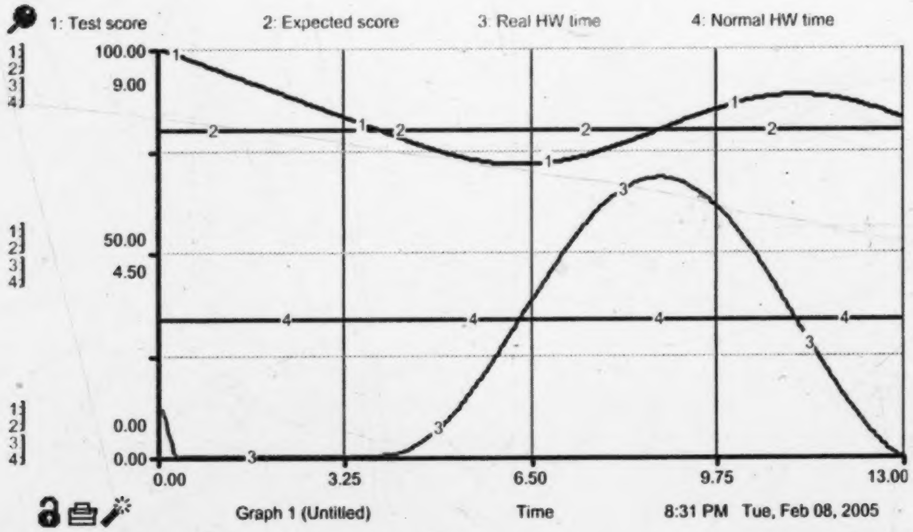


Figure 6.6: Behaviour of the variables "Test score" and "homework time" in the model based on the loop "Studying"

While each variable represented on the STELLA output may have its own scale; in this case the scales of the "Real HW Time" and "Normal HW Time" are the same, which is indicated by the brackets on the left side of the diagram (the same is true for the variables "Test score" and "Expected score").

As the semester begins, the student's actual knowledge is assumed to be equal to the required knowledge (which is zero at the beginning). Since the student was expecting a mark of 80, the student will reduce the amount of hours per week spent on homework. The homework time will gradually drop to zero and will stay at zero until the test score drops below the expected test score (approximately 4 weeks into the semester). At that time, the student will increase the amount of time per week spent on homework, but the test score will continue to fall until the actual homework time reaches the required homework time (approximately 6.5 weeks into the semester). Since the test score is still below the expected score, the student will continue increasing the amount of homework time per week until the test score reaches the expected level (approximately 8.7 week into the semester). At that time, the student will reduce the number of hours per week spent on homework, but as long as they stays above the normal amount of homework time (until approximately 11 weeks into the semester), the test score will continue to rise. When the homework time drops below the normal time, the test score will start to decline.

6.3.3.4. Model testing

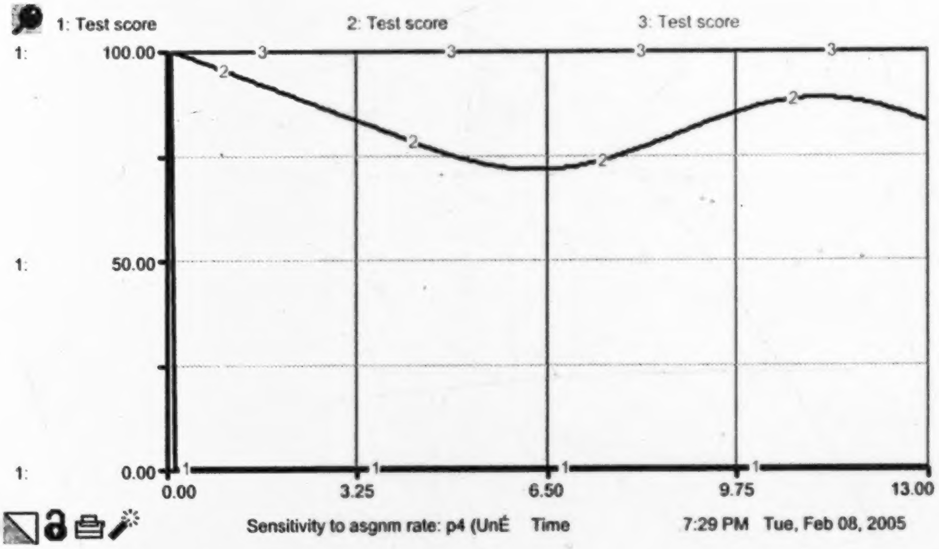
Several tests were conducted on the model to examine its quality. One of the tests requires the introduction of extreme values for some of the model parameters. The model should exhibit reasonable behaviour even under the extreme (unrealistic) specifications. The model behaviour was tested under these extreme values of the assignment rate:

1. extremely high rate (10,000 assignments per week)
2. normal rate (1 assignment per week)
3. zero rate of assignments per week

The behaviour of the variable "Test score" in a SD model based on the loop "Studying" under these conditions is illustrated in Figure 6.7.

Under the assignment load that is artificially extremely high, the required number of hours of homework per week will also be extremely high. A student will be consistently behind in the amount of knowledge required for any mark above zero (since, in the model, it is assumed that knowledge gain comes from the self-studying only), and, consequently, the test score will be zero throughout the course.

When no assignments are required, the required knowledge gain (that comes from working on assignments) will be zero. The actual gain will be zero as well, since knowledge gain occurs from self-studying only. The relative knowledge, therefore, will be equal to the required knowledge (and both will be equal to zero), and a student will receive test score of 100.



1: Test score – behaviour of test score under scenario 1

2: Test score – behaviour of test score under scenario 2

3: Test score – behaviour of test score under scenario 3

Figure 6.7: Behaviour of "Test score" under extreme values of the assignment rate

The system can achieve balance when the expected test score is 100, and the initial actual homework time (stock "Real HW time") is equal to the required homework time (converter "Normal HW time") (see Figure 6.8). The zero difference between the required homework time and the actual homework will create a zero knowledge gain (which is a function of the homework time difference). The zero knowledge gain rate means that the relative knowledge (e.g., the difference between the required and the actual levels of knowledge) will be zero or, in other words, that the student will be gaining amount of knowledge just sufficient to maintain the desired grade.

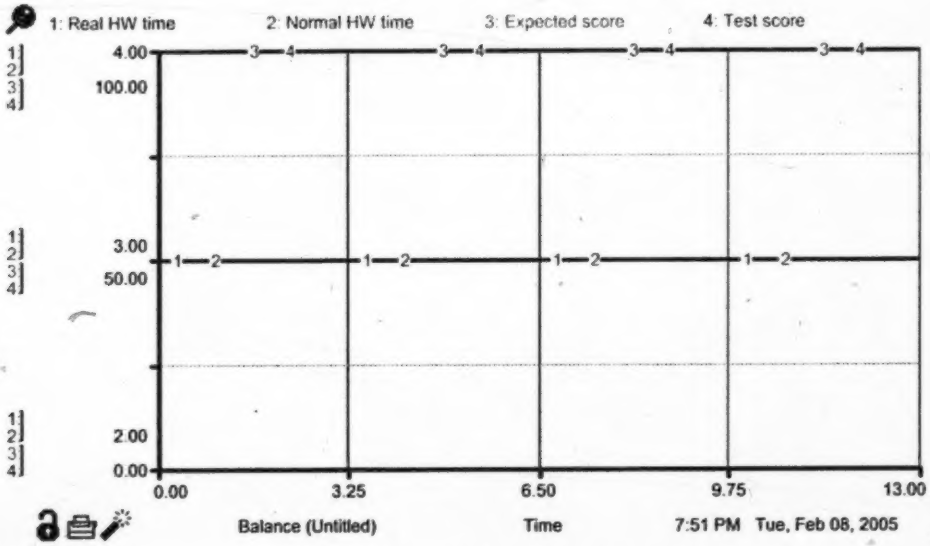


Figure 6.8: Balance state of the model based on the loop "Studying"

6.3.4. Complete STELLA model

By using the causal loop diagram describing the whole system (Figure 6.4), and a stock-and-flow model based on the loop "Studying" (Section 6.3.3.), we can construct a complete stock-and-flow model by using STELLA software. The stock-and-flow map is presented in Figure 6.9.

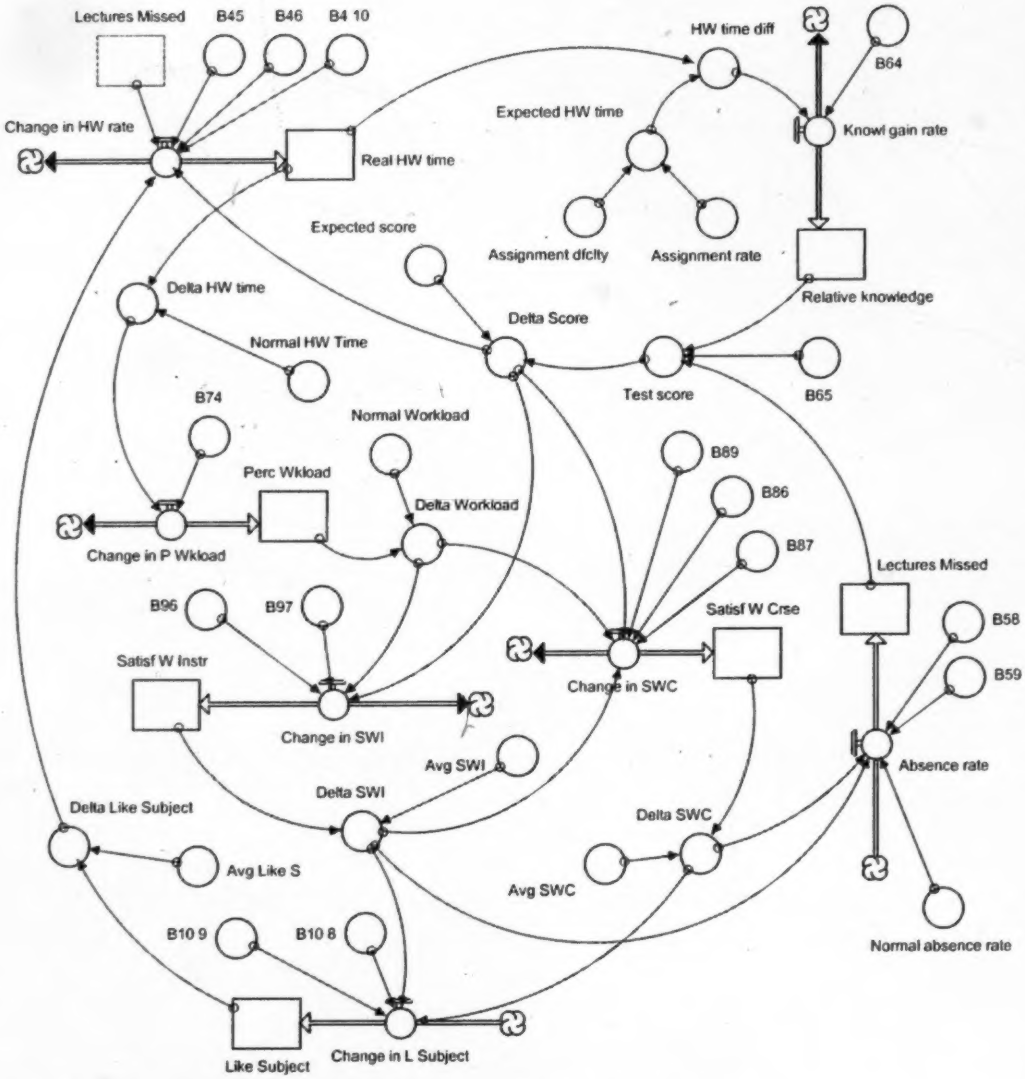

Lectures Missed
B45
B46
B4 10
HW time diff
B64
Change in HW rate
Expected HW time
Knowl gain rate
Real HW time
Expected score
Assignment dfclty
Assignment rate
Delta HW time
Relative knowledge
Delta Score
Normal HW Time
Test score
B65
B74
Normal Workload
B89
Perc Wkload
B86
Delta Workload
B87
Change in P Wkload
Lectures Missed
B96
B97
Satisf W Crse
B58
Satisf W Instr
Change in SWC
B59
Change in SWI
Avg SWI
Absence rate
Delta SWI
Delta Like Subject
Delta SWC
Avg Like S
Avg SWC
B10 9
B10 8
Normal absence rate
Like Subject
Change in L Subject

Figure 6.9: Stock-and-flow map for the complete STELLA model

6.3.5. Model equations and starting values

In assigning values to the constants "Average satisfaction with instructor," "Average satisfaction with course," "Normal workload," "Normal homework time," and

"Average attitude toward the subject," an "average" student was modeled. The normal absence rate of 0.5 represented one lecture missed every two weeks. The normal homework time equal to 3 hrs per week represented an average amount of time spent on a homework assignment in an engineering course. The values of average satisfaction and attitude were given a value of 3 on the students' questionnaire (see Table 5.3), and, therefore, were set at a value of 3 in the STELLA model. The assignment rate of one per week and the assignment difficulty of 3 hours of homework per assignment represented the average parameters for the course. The complete set of model equations is presented in Appendix XVI.

6.3.6. Behaviour of the complete model

The behaviour of the model's variables under the "average student" set of parameters is illustrated in Figures 6.10 and 6.11.

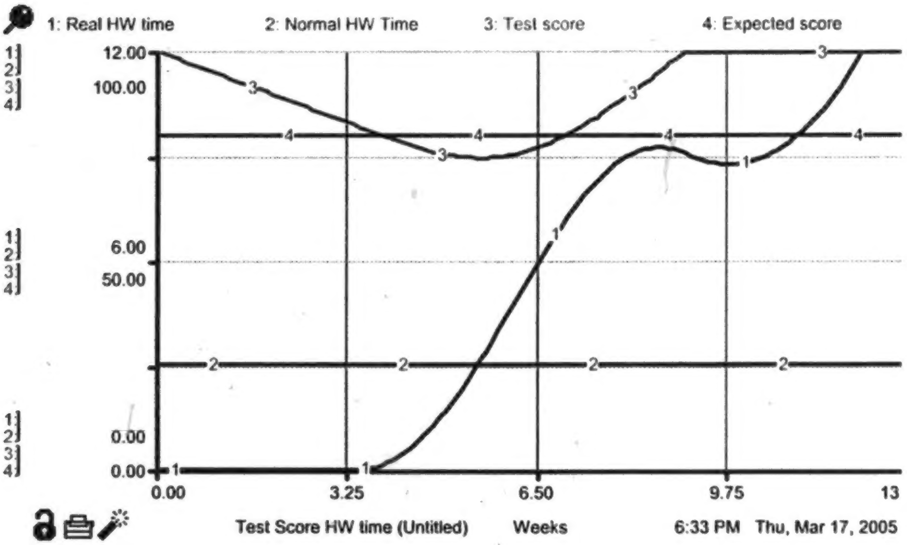


Figure 6.10: Behaviour of the variables "Test score" and "Homework time" in the complete model

The pattern of the behaviour of the variables "Test score" and "Homework time" in the complete system (see Figure 6.10) is similar to the variables' behaviour in the system based on the loop "Studying" (see Figure 6.6) until approximately 7 weeks into the semester. Similarly to the behaviour of the model based on the loop "Studying," the test score drops below the expected score at approximately 3.8 weeks into the semester, and the homework time starts to rise. The test score in the complete model, however, reaches the expected score sooner (7 weeks vs. 8.7 weeks), and, unlike in the model based on the loop "Studying," the homework time in the complete model does not fall below the normal homework time, but keeps rising.

This behaviour of the homework time in the complete model is caused by the reinforcing loops "Attitude Effect I" and "Attitude Effect II". The behaviour of the variables "Satisfaction with the course" (Satisf W Crse), "Satisfaction with the instructor" (Satisf W Instr in Figures), "Attitude toward the subject" (Like Subject), and "Perceived Workload" (Perc Wkload) is illustrated in Figure 6.11:

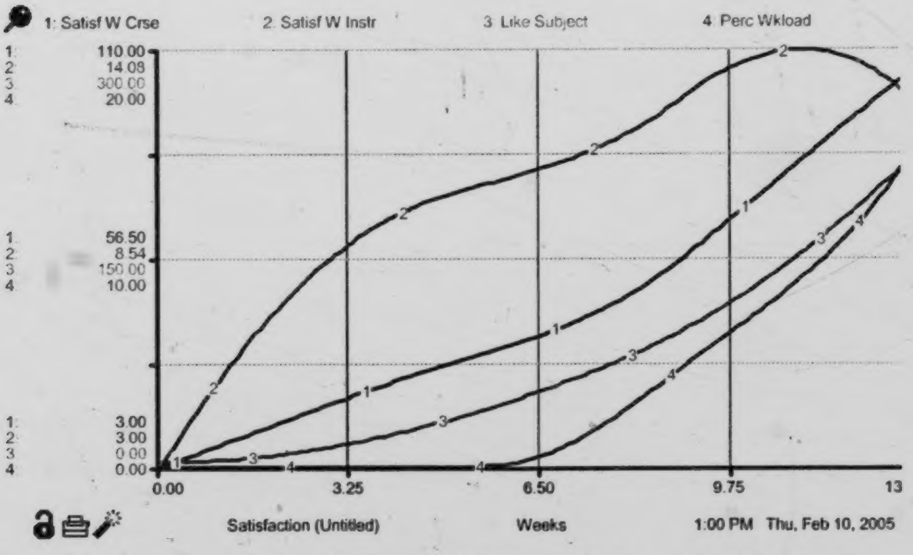


Figure 6.11: Behaviour of the variables "Satisfaction with the course," "Attitude toward the subject," and "Perceived workload" in the complete model

Even though the perceived workload increases with the increase in homework time, the actions of the balancing loops “Workload Effect I” and “Workload Effect II” are not sufficient to counteract the actions of the reinforcing loops.

Extending the model horizon to 40 weeks reveals a shift in the loop dominance (Figure 6.12):

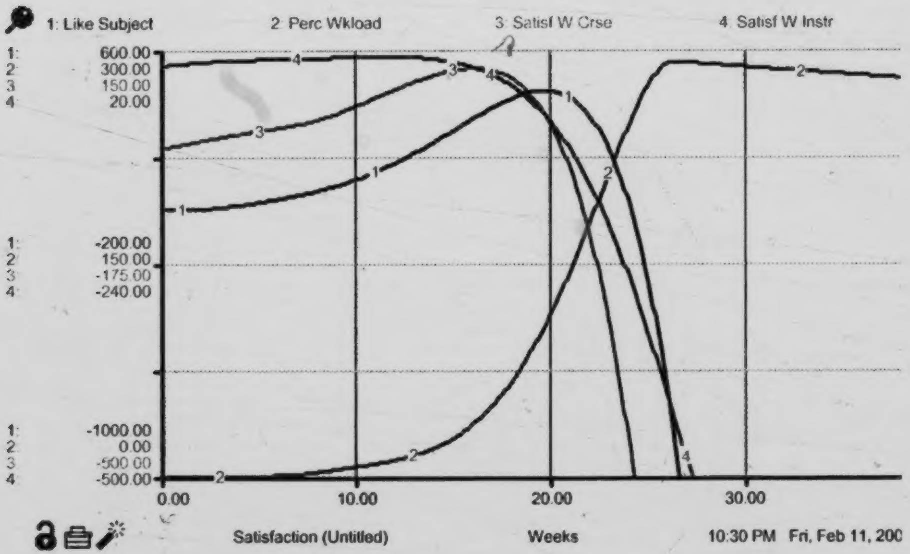


Figure 6.12: Shift in loop dominance

As the perceived workload rises with the increase in homework time, the dominance shifts from the reinforcing loops “Attitude Effect I” and “Attitude Effect II” to the balancing loops “Workload Effect I” and “Workload Effect”. A high level of workload drives down satisfaction with course and with the instructor, and this decline, in turn, drives down the attitude toward the subject. When the measured attitude toward the subject starts decreasing (around approximately 22.6 weeks), the homework time starts decreasing, and falls to zero by 26.4 weeks (Figure 6.13):

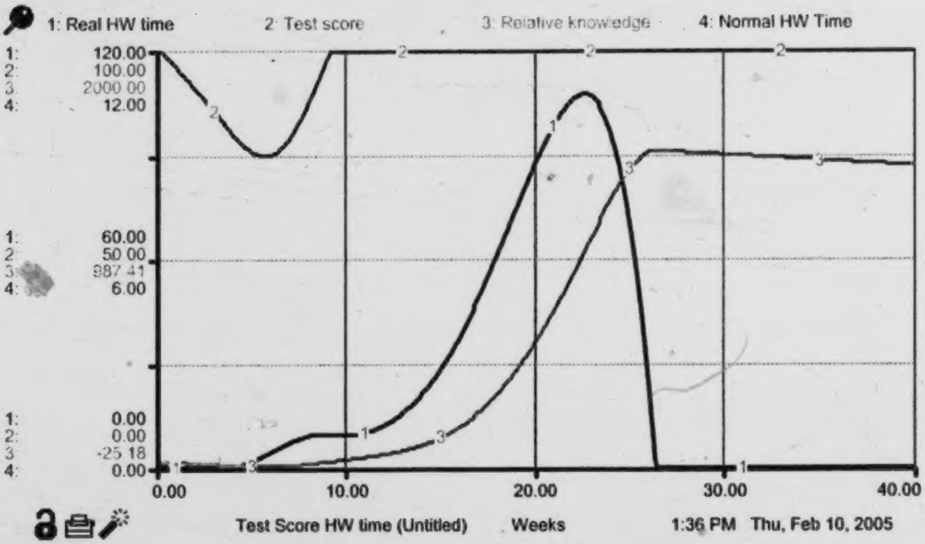


Figure 6.13: Behaviour of the complete model over a 40-week time horizon

As the homework time falls to zero, the relative knowledge starts decreasing. The test score stays at 100, however, since over the period of time when the homework time stayed above the required homework time (which was 3 hours), the student accumulated a significant amount of "extra" knowledge above the required level (around 1,500 units at time 26.4 weeks). Since the relative knowledge is reduced by the difference in homework time, the difference of "–3" (real HW time – expected HW time) drains the stock of relative knowledge very slowly.

The test score eventually falls to zero. The relative knowledge and the perceived workload do decline as the homework time drops to zero, but the high accumulated stock of the perceived workload increases the stock of the lectures missed. Eventually, the number of lectures missed reaches a level where it brings the test score down to zero. This result occurs at approximately 67 weeks (see Figure 6.14):

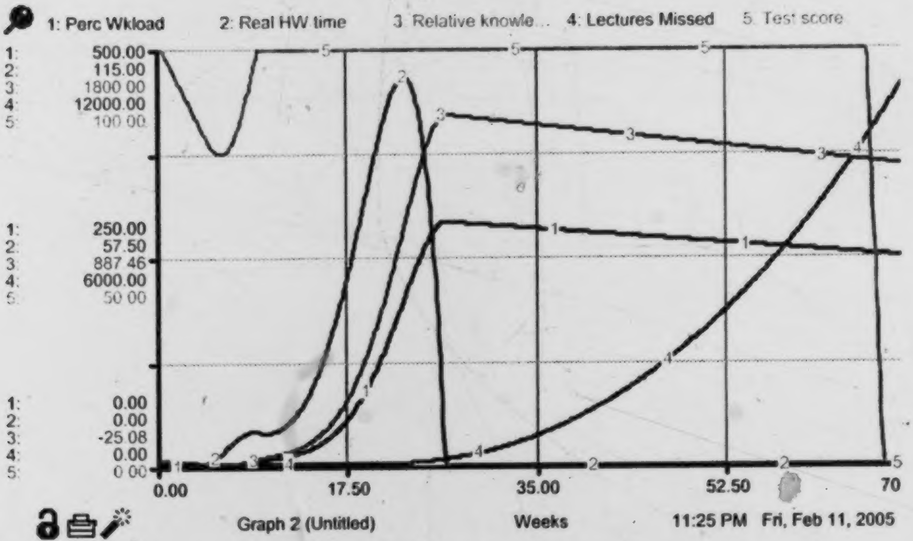


Figure 6.14: Behaviour of the complete model over a 70-week time horizon

6.3.7. Model testing

The model was tested by analyzing the sensitivity of the variable "Test score" to the changes in the model's parameters and by evaluating the model's behaviour under extreme conditions.

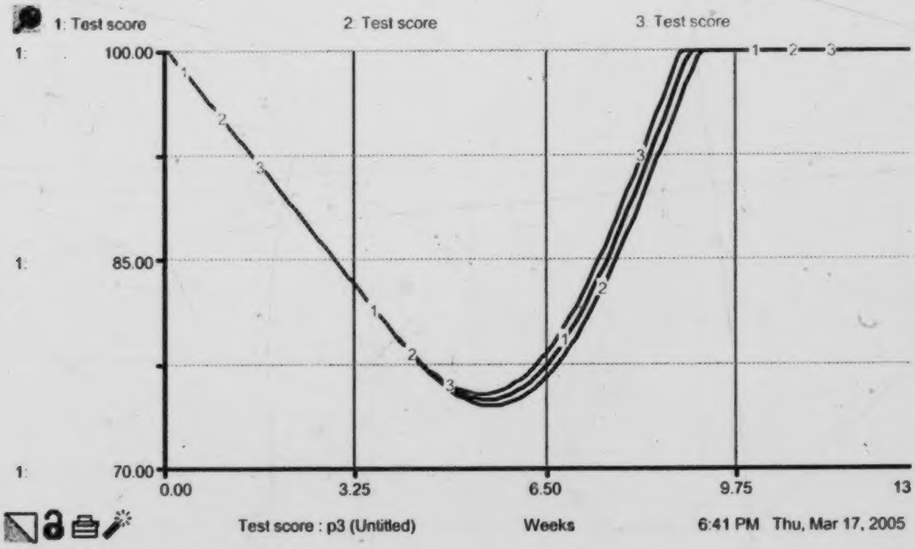
6.3.7.1. Satisfaction with the course and the instructor, and attitude toward subject

The sensitivity of the test score to the changes in students' average satisfaction with the course, average satisfaction with the instructor, and average attitude toward the subject was tested by changing one variable at a time. Each variable ("Satisfaction with the course," "Satisfaction with the instructor," and "Attitude toward the subject") had three levels: "Normal" (value set at 3), "High standards" (value set at 5), and "Low standards" (value set at 1). A low (high) standard characterizes a situation when an incoming student has a low (high) average satisfaction with either a similar course, a similar instructor, or a similar subject. Table 6.1 below specifies the parameter values corresponding to each scenario:

Table 6.1: Parameters for test score sensitivity analysis scenarios

Scenario	Scenario	Parameter value
1	"Average"	3
2	"High standard"	5
3	"Low standard"	1

The model's time horizon was set at 13 weeks. The graphs below (see Figures 6.15 – 6.17) illustrate the behaviour of the variable "Test score" under each scenario.

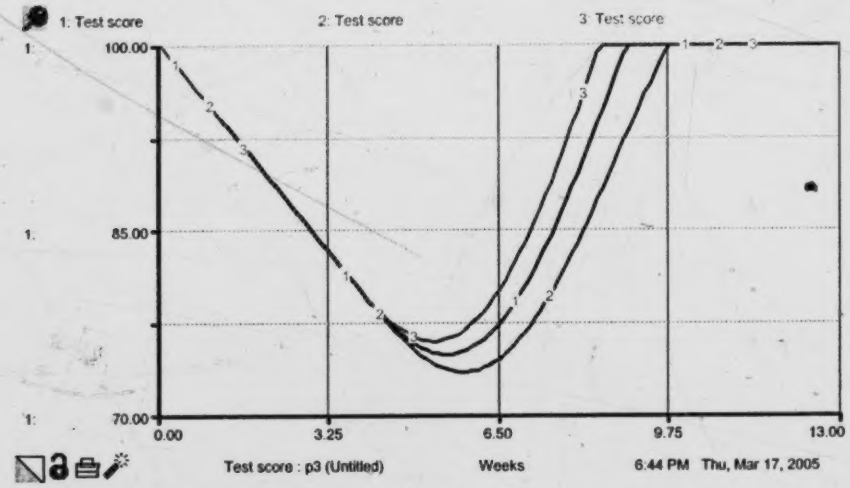


1: Test score – behaviour of test score under scenario 1

2: Test score – behaviour of test score under scenario 2

3: Test score – behaviour of test score under scenario 3

Figure 6.15: Sensitivity of the variable "Test score" to the changes in the variable "Satisfaction with the course"

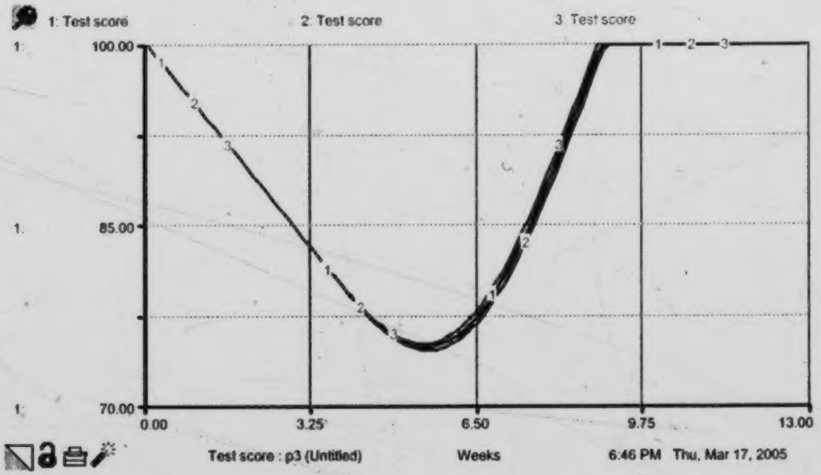


1: Test score – behaviour of test score under scenario 1

2: Test score – behaviour of test score under scenario 2

3: Test score – behaviour of test score under scenario 3

Figure 6.16: Sensitivity of the variable "Test score" to the changes in the variable "Satisfaction with the instructor"



1: Test score – behaviour of test score under scenario 1

2: Test score – behaviour of test score under scenario 2

3: Test score – behaviour of test score under scenario 3

Figure 6.17: Sensitivity of the variable "Test score" to the changes in the variable "Attitude toward the subject"

All three graphs illustrate that under a "High standard" scenario (#2), the test score falls relatively lower below the expected score than that of an "average" or "normal" student. When a student has low expectations about the course, instructor, and subject, a perceived quality of the course, instructor, or subject higher than the "expected low" would lead to high levels of satisfaction and attitudes. The test score of such a "low standards" student does not drop as far below the expected score as that of an "average" or "high standards" student.

However, the overall difference in the test scores "average," "high standard," and "low standard" students is not significant. The largest difference (2.54) between the lowest values of the test score was attributed to the change in the average satisfaction with the instructor (Figure 6.16). The low sensitivity of the test score to the changes in the average satisfaction with the course, average satisfaction with the instructor, and in the average attitude toward the subject resulted because each feedback loop involving the variables "Satisfaction with course," "Satisfaction with instructor," and "Attitude toward subject" contained at least one statistically non-significant causal link (see Figure 6.4).

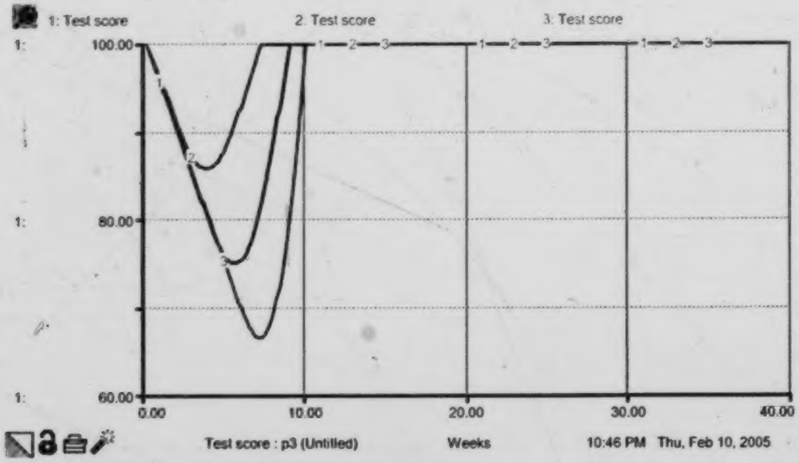
6.3.7.2. Sensitivity analysis: the "good" versus the "bad" student

One of the sensitivity analysis tests was to simulate the performance of a "good student" versus the performance of a "bad student." A "good" student would be one who has high expectations of the subject, instructor, and course; normally spends a reasonable amount of time studying; and attends all lectures. A "bad" student would be one who has low expectations of the subject, instructor, and course; normally does not spend any time doing homework; and misses most of the lectures. The test scores of the "good" and "bad" students were compared with the test score of an "average" student.

The model's parameters for the "good" and "bad" students are presented in Table 6.2. Except for the variables listed in Table 6.2, all other model parameters were the same for both the "good" and "bad" students. The behaviour of the variable "Test score" is illustrated in Figure 6.18. Sensitivity analysis is carried out over 40 weeks to best capture the dynamics of the system.

Table 6.2: "Good" versus "bad" student model parameters

Parameter	"Average" student (Scenario #1)	"Good" student (Scenario #2)	"Bad" student (Scenario #3)
Expected test score	80	95	60
Normal homework time, hrs/week	3	5	0
Normal workload	3	5	1
Average satisfaction with the course	3	5	1
Average satisfaction with the instructor	3	5	1
Average attitude toward the subject	3	5	1
Normal absence rate, lectures/week	0.5	0	2



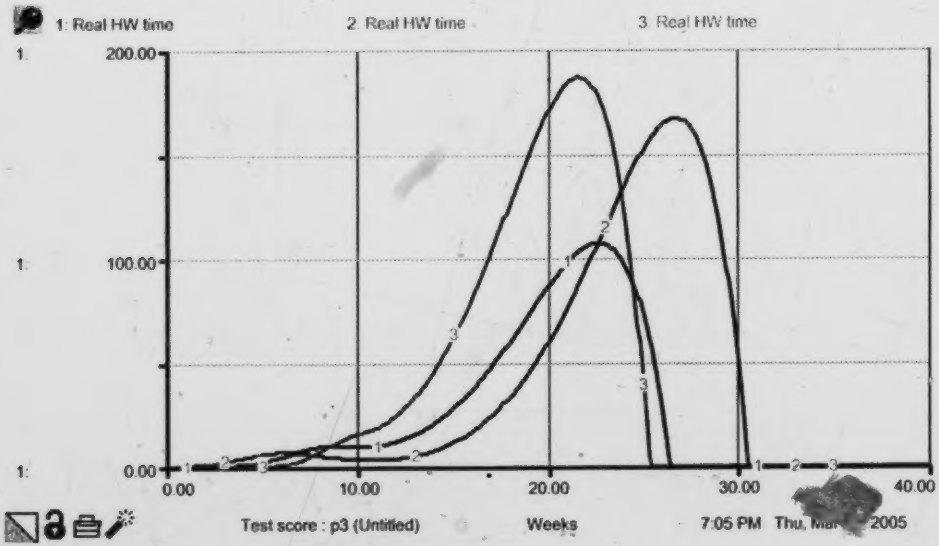
1: Test score – behaviour of test score under scenario 1

2: Test score – behaviour of test score under scenario 2

3: Test score – behaviour of test score under scenario 3

Figure 6.18: Behaviour of the variable "Test score" under the various student scenarios

The peak homework time is lower for the "good" student, and higher for the "bad" student (Figure 6.19). Also, a "good" student aims for a higher mark and does not wait long to start increasing the time spent on homework (homework time starts increasing at time 1.5 weeks). A "bad" student may wait longer before starting to spend more time on homework, since the student aims for a lower mark (homework time starts increasing at 5.6 weeks) (Figure 6.20).

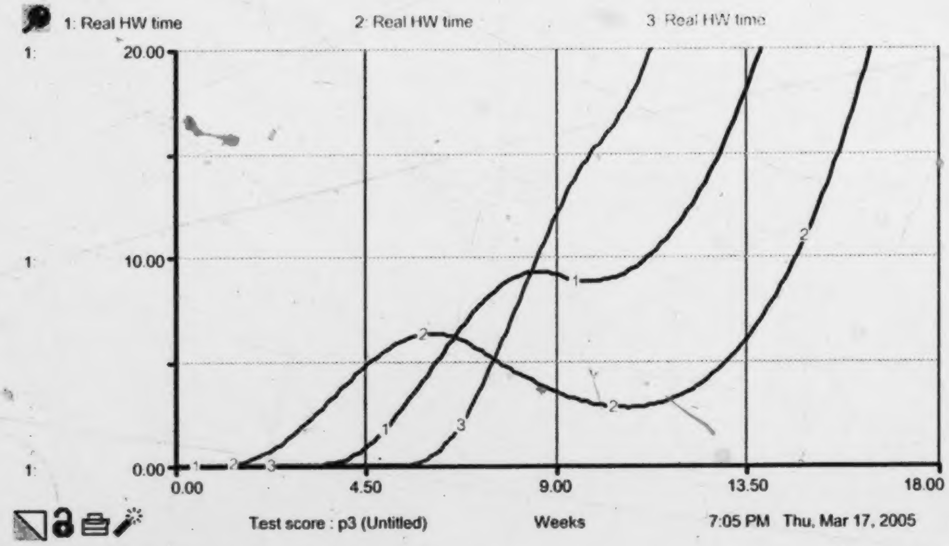


1: Real HW time – behaviour of real HW time under scenario 1

2: Real HW time – behaviour of real HW time under scenario 2

3: Real HW time – behaviour of real HW time under scenario 3

Figure 6.19: Peak homework time



1: Real HW time – behaviour of real HW time under scenario 1
2: Real HW time – behaviour of real HW time under scenario 2
3: Real HW time – behaviour of real HW time under scenario 3

Figure 6.20: "Good" versus "bad" student: timing of increase in homework time

6.3.7.3. Extreme values

The model's performance under extreme conditions must remain realistic. The output of the decision rule has to be reasonable even under input values not observed in the real world (Sterman 2000).

This model assumed that student's relative knowledge depended on homework time only (obviously, a simplification). Therefore, if the homework time was zero, we would expect to see relative knowledge becoming increasingly negative, and the test score to fall to zero. The behaviour of the variables "Test score" and "Perceived Workload" (Perc Wkload) under zero homework hours is illustrated in Figure 6.21. In this figure, "Zero HW" is the line illustrating zero hours spent on homework.

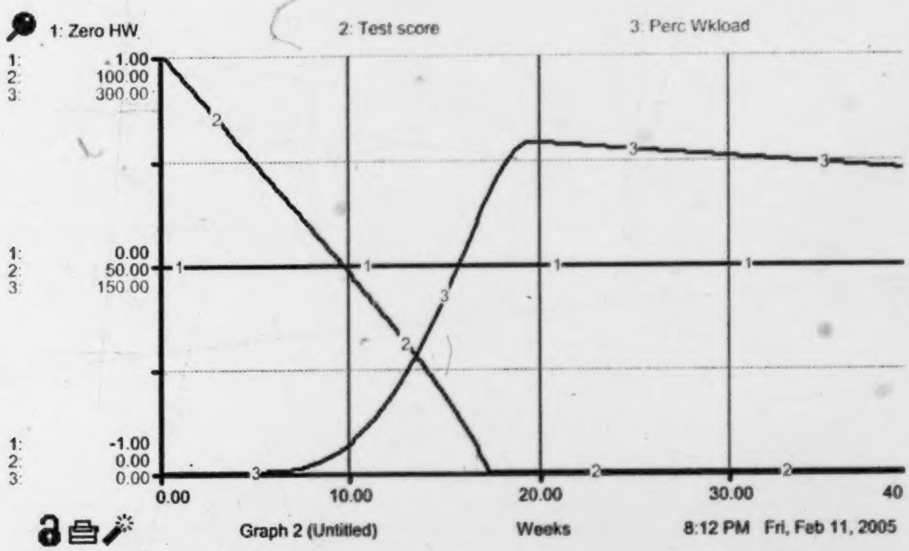
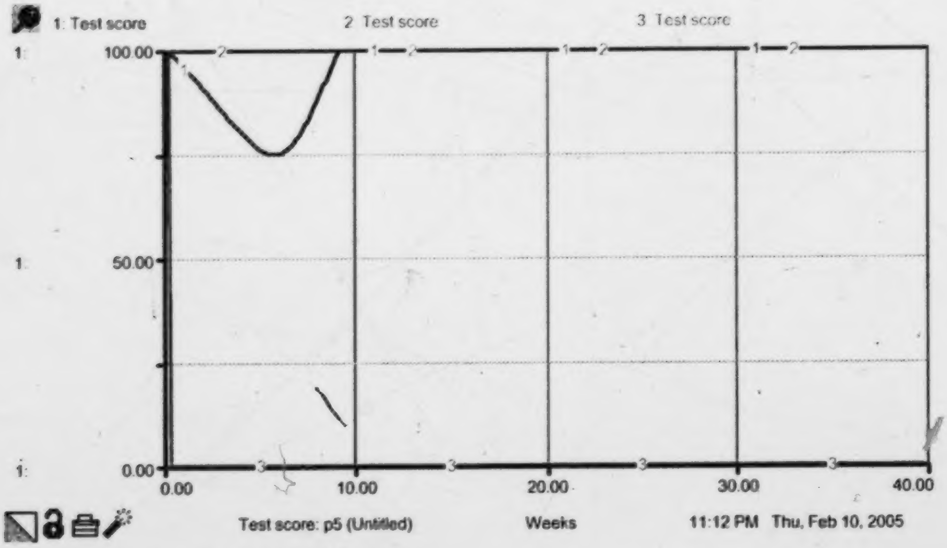


Figure 6.21: Behaviour of the variable "Test score" under the zero homework hours scenario

The behaviour of the variable "Test score" was also tested by the extreme values of the variable "Assignment rate." Compared to the normal rate of 1 assignment per week, values of 0 and of 10,000 were used to evaluate the model's validity (similarly to the testing of the model based on the loop "Studying," Section 6.3.3.4.). The behaviour of the variable "test score" under the normal (scenario #1), zero (scenario #2), and extremely high (scenario #3) rates of assignments is illustrated in Figure 6.22:



1: Test score – behaviour of test score under scenario 1

2: Test score – behaviour of test score under scenario 2

3: Test score – behaviour of test score under scenario 3

Figure 6.22: Behaviour of the variable "Test score" in the complete model under the extreme values of the variable "Assignment rate"

The behaviour of the complete model is similar to the behaviour of the model based on the loop "Studying." When no homework is assigned during a week, a student will maintain a 100 mark. If the number of assignments is extremely high, the student's mark will drop to 0 immediately (technically, after the first model iteration).

6.3.7.4. Variables' ranges and scales

Figures 6.11 – 6.22, reveal that some of the model's variables take on "unrealistic" values. For example, the attitude toward the subject falls to negative 1,000 in Figure 6.12, homework time per week reaches above 100 hours in Figure 6.13, and relative knowledge rises to about 1,500 units in Figure 6.13. Does these results invalidate the model?

The answer is no for two reasons.

Firstly, while physical parameters such as weight or speed can be measured and quantifies, "soft" variables like satisfaction and attitude are virtually impossible to measure, and can only be quantified (Richmond et al. 2000). All measurement scales for unobserved variables are fictions whose scales are set arbitrarily by modelers (Hayduk 1987).

Secondly, the model's validity is tested relative to the model's purpose. The purpose of this effort was not to create a perfect model of a classroom system, for such a model would have been just as complex as the classroom itself. The goal of the modeling effort was to examine the role of the soft variables in the system's dynamics and not to predict the numeric magnitude of a particular attitude factor.

This model exhibited reasonable behaviour. The patterns of the changes in the variables could be explained by examining the underlying causal structures. When the model was expected to produce a zero test score, it did so; when the expected outcome was a constant test score of 100, the outcome confirmed this expectation.

Therefore, one can argue that this model passed the validation tests and can now be used to examine the effect of policy changes on teaching an undergraduate engineering management course.

6.3.8. Policy testing

The essential goal of a modeling effort is to resolve the issue that caused the creation of the model in the first place (Sterman 2000). One of the purposes of creating a model of classroom performance was to find policies that are effective in increasing students' knowledge and improving their attitude toward the subject. Effective policies are "leverage points": a small change to an input factor produces a significant change in the system's behaviour (Richmond et al. 2000).

An effective policy must be realistic, and be under the decision-maker's influence (Richmond et al. 2000). For example, a student's language background or age influences the student's level of knowledge and attitude (as was indicated by the structural coefficients in the SEM model), but will the policy aimed at influencing those parameters be effective? The answer is "Most likely not."

The variables that are under an instructor's direct influence in this model are the assignment difficulty and rate. Another variable that can be indirectly influenced is the lecture attendance. An instructor might *compel* students to attend lectures by including on the midterm test the questions that are discussed during lectures only (this situation actually happened).

An instructor might *encourage* students to attend lectures by inspiring their interest through the use of innovative lecturing techniques, creativity, challenging discussions, relevant examples and applications from real life.

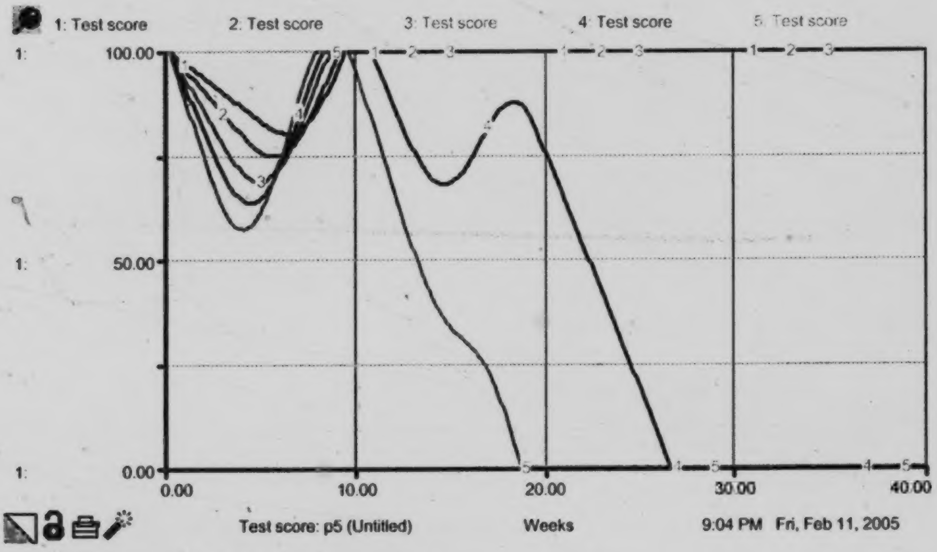
6.3.8.1. Assignment policy

It was decided to test the sensitivity of the variables "Test score" and "Attitude toward the subject" by changing the assignment rate and assignment difficulty, and by using a combination of both (see Table 6.3). The rest of the parameters were set at the "average student" level.

Table 6.3: Assignment policies

Parameter	Scenario #1	Scenario #2 (baseline)	Scenario #3	Scenario #4	Scenario #5
Assignment rate, #/week	1	1	1.5	2	2
Assignment difficulty, hrs/assignment	2	3	3	3	4
Effective required HW time, hrs/week	*2*	*3*	*4.5*	*6*	*8*

Behaviour of the variables "Test score" and "Attitude toward subject" is illustrated in Figures 6.23 and 6.24.



1: Test score – behaviour of test score under scenario 1

2: Test score – behaviour of test score under scenario 2

3: Test score – behaviour of test score under scenario 3

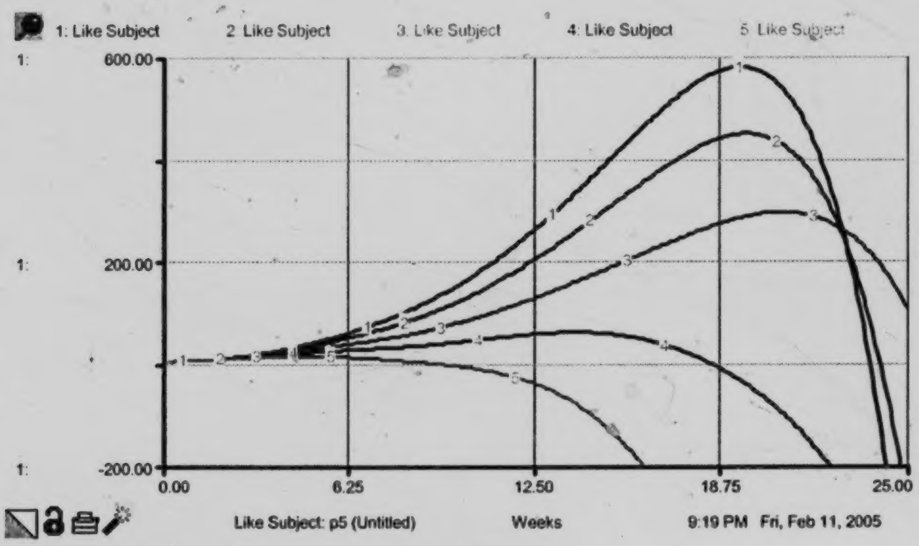
4: Test score – behaviour of test score under scenario 4

5: Test score – behaviour of test score under scenario 5

Figure 6.23: Behaviour of the variable "Test score" under various assignment policies

Figure 6.23 reveals that the lighter the assignment workload during the course, the easier a student can achieve and maintain a mark of 100. At some workload level, however, a threshold is passed, and the student is not able to maintain a 100 mark. This situation should be familiar to instructors: the higher the standard, the fewer students will be able to achieve it, and the more difficult achieving the standard will be.

Figure 6.24 illustrates the behaviour of the variable "Attitude toward the subject" (Like Subject in Figure 6.24) over the period of 25 weeks.



1: Like subject – behaviour of attitude toward subject under scenario 1
2: Like subject – behaviour of attitude toward subject under scenario 2
3: Like subject – behaviour of attitude toward subject under scenario 3
4: Like subject – behaviour of attitude toward subject under scenario 4
5: Like subject – behaviour of attitude toward subject under scenario 5

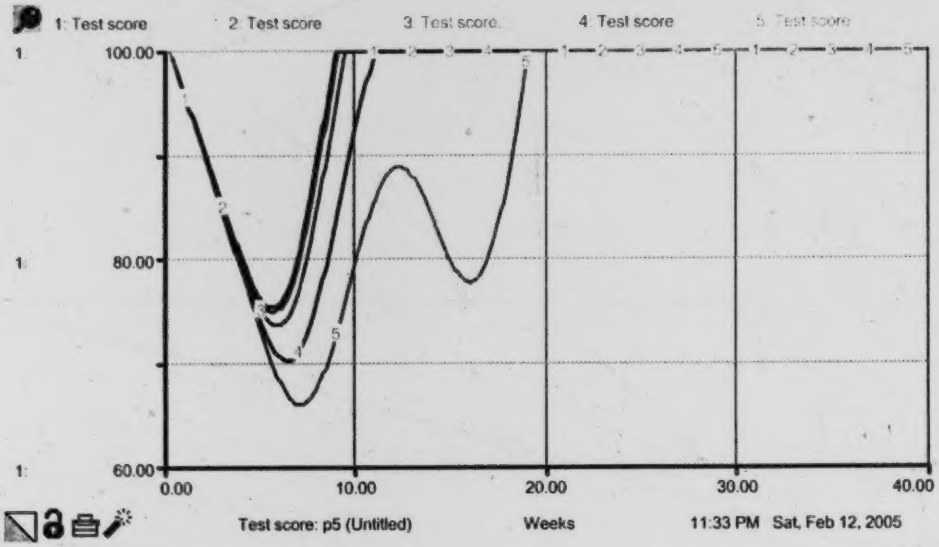
Figure 6.24: Behaviour of the variable "Attitude toward the subject" under various assignment policies

Firstly, with a lighter workload, attitude reaches a higher maximum value (583 during scenario #1 versus 12 during scenario #5). Secondly, loop dominance shifts from the positive to the negative loops sooner with the highest workload (week 4.7). The shift in dominance occurs at the latest time for a scenario with the average workload (21 weeks, scenario #3). Thirdly, the shift in dominance is smooth with the highest workload (scenario #5), and is abrupt with the lowest (scenario #1). Compared to the system with a heavier workload, the attitude toward the subject in the system with a lower workload rises to a higher, and falls to a lower value.

6.3.8.2. Attendance policy

The sensitivity of the variables "Test score" and "Attitude toward subject" to the lecture attendance was tested by simulating the normal absence rates of 0 lectures per week (scenario #1), 0.5 (scenario #2), 1 (scenario #3), 2 (scenario #4), and 3 lectures per week (scenario #5). The behaviour of the test score under these different absence scenarios is illustrated in Figure 6.25.

Figure 6.25 indicates that a low-to-moderate normal absence rate does not significantly affect student performance (scenarios 1-3). When the normal absence rate increases to 2 per week and 3 per week, achieving and maintaining mark of 100 becomes more difficult for a student.



1: Test score – behaviour of test score under scenario 1

2: Test score – behaviour of test score under scenario 2

3: Test score – behaviour of test score under scenario 3

4: Test score – behaviour of test score under scenario 4

5: Test score – behaviour of test score under scenario 5

Figure 6.25: Behaviour of the variable "Test score" under various attendance scenarios

Under scenario 5, the real absence rate drops to 2.08 lectures per week, since the real absence rate depends on satisfaction with the course (Satisf W Crge in Figure 6.26) and satisfaction with the instructor (Satisf W Instr in Figure 6.26), besides depending on the normal absence rate. As long as satisfaction with the course and instructor increases, the absence rate decreases. Satisfaction, in turn, depends on the level of the perceived workload (Perc Wkload in Figure 6.26). When the perceived workload starts to increase (when the dominance of the loops changes), the absence rate starts to rise as well (see Figure 6.26):

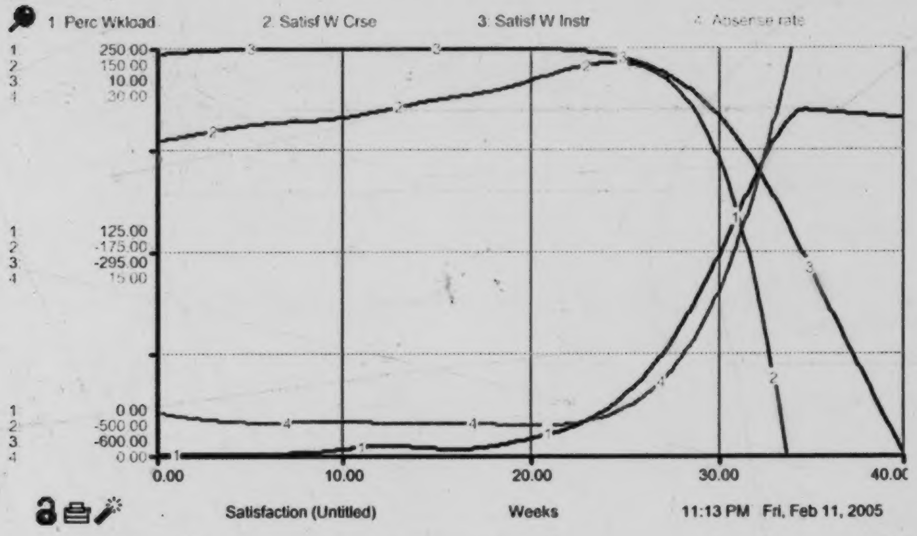
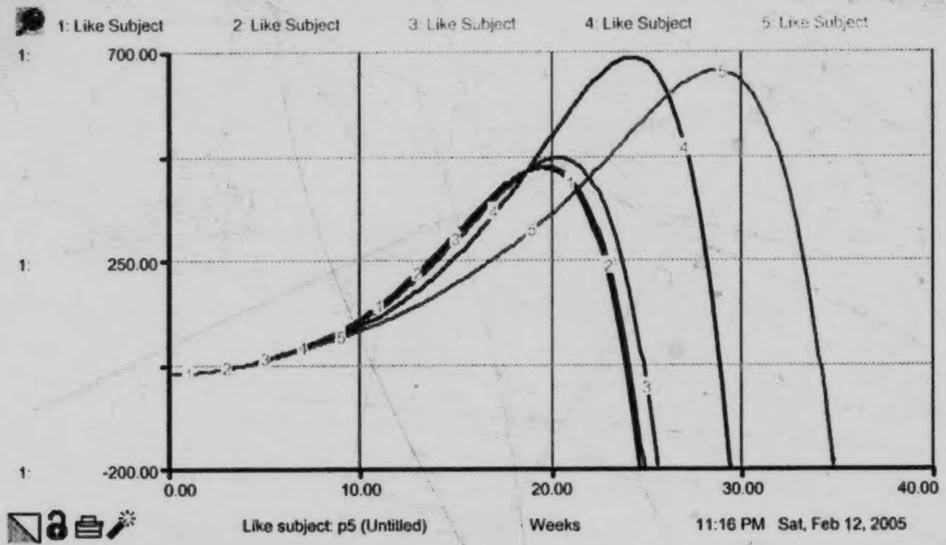


Figure 6.26: Rise in absence rate, scenario #5

While the stock of the perceived workload starts to decrease at approximately 34 weeks (when the homework time drops to zero), the high accumulated level of the perceived workload drives down the values of the satisfaction, and this decrease, in turn, drives the absence rate up.

The behaviour of the variable "Attitude toward the subject" under different attendance scenarios is illustrated in Figure 6.27. The attitude toward the subject rises to the highest level under scenario #4 (690 units) and falls to the least negative level at the end of simulation under scenario #5 (-3,800 units). The patterns of the attitude toward the subject under scenarios 1-3 are similar: the attitude rises to about the same level (450 to 470 units) at approximately the same time (19-20 weeks) and falls to about the same level (negative 30,000 to negative 26,700) at the end of simulation. At week 13, the attitude toward the subject is positive and increasing under all scenarios.



1: Like subject – behaviour of attitude toward subject under scenario 1

2: Like subject – behaviour of attitude toward subject under scenario 2

3: Like subject – behaviour of attitude toward subject under scenario 3

4: Like subject – behaviour of attitude toward subject under scenario 4

5: Like subject – behaviour of attitude toward subject under scenario 5

Figure 6.27: Behaviour of the variable "Attitude toward the subject" under various attendance scenarios

6.4. Summary

The SD model presented in this chapter illustrated that a classroom is a complex, dynamic, time-sensitive environment. To anyone working in the field of education this finding is not a revelation. However, a significant amount of time and effort being currently spent by researchers on understanding the classroom system is an indication that the knowledge of the nature and behaviour of the system is far from complete and comprehensive.

One can gain insight into the nature of the relationships in a classroom system by using a statistical modeling technique such as SEM, but if the system is to be modified, statistical models are of limited use since they do not include time as a variable. Mathematical optimization tools do not work when an analytical description of a model is not available, as is often a case with social systems.

This chapter illustrated how model formulation can be put on a solid basis by using estimates of structural relationships obtained by first modeling a classroom system in SEM. The SD model demonstrated that even all-linear relationships embedded in a feedback structure produce complex dynamic behaviour.

The SD model was used to simulate the behaviour of particular system variables – "Test score" and "Attitude toward the subject" – under different homework workload and attendance scenarios. The insight gained from the simulations will be used in modifying the course content and delivery. This process is the subject of Chapter 7.

Chapter 7: Course Improvement

7.1. Introduction

A principle central to the quality management philosophy is that quality improvement should be, first, a systematic, rather than an ad hoc effort, and, second, that the effort should be continuous, rather than one-time. This chapter will illustrate how a continuous systematic effort will be applied to improve such performance indicators as students' satisfaction with the course, satisfaction with the instructor, and attitude toward the subject.

7.2. Continuous improvement

The only changes introduced into the educational system should be the changes producing improvement, and the improvement should be continuous (Jenkins 2003). Jenkins (2003) suggested the following sequence for the continuous improvement process in the K-12 educational system:

1. Gathering data;
2. Constructing graphs;
3. Gaining insight from studying graphs;
4. Testing hypotheses;
5. Increasing knowledge from hypotheses testing.

This sequence closely resembles the well-known a Plan-Do-Check-Act (PDCA) approach, introduced by Dr. W. Edwards Deming (1986). The ISO 9001 (2000) standard illustrates how the PDCA approach can be used to achieve improvement in the quality of products and services. While different interpretations of each stage of the PDCA approach are available (see, for example, Montgomery 1997a, or ISO 9001 (2000)), the following interpretation is suggested for use at the classroom level:

- Plan: design policies aimed at improving selected course processes, outputs, or outcomes;
- Do: implement changes in the course delivery and content;
- Check: collect data to evaluate whether the changes produced the desired result;
- Act: introduce adjustments as required.

7.3. Application

7.3.1. Plan

The course quality improvement efforts in this research were directed at improving three educational outcomes: students' satisfaction with the course, satisfaction with the instructor, and attitude toward the subject. As well, this research was also directed at improving the process of knowledge transfer.

The ideas for improving the process and outcome came from the analysis of available quantitative and qualitative data. The lessons learned from the application of MBKP and SPC in 2002-2003 were used to plan improvements to the process of knowledge transfer. The outputs of the SEM and SD models were used to plan improvements that would lead to better educational outcomes.

Outputs of the SPC, SEM, and SD models were the sources of the quantitative data. Customer feedback is the source of qualitative data about a system's products and is one of the most important sources of information about possible areas for improvement. The questionnaire described in Chapter 5 not only provided data for a statistical model of the system, but also contained a section for students to provide their suggestions and recommendations. Even though such feedback is provided in qualitative rather than quantitative form, it may actually contain an even greater amount of useful information than quantitative data and definitely should not be ignored (Luna-Reyes and Anderson 2003, Sterman 2000). Out of 384 questionnaires collected during Fall 2003 – Winter 2004 semesters, 71 (18.5%) contained written comments and suggestions. The prevailing themes were complaints about the number of assignments in the course, and the length of the midterm exam (the survey was conducted one week after the first midterm exam). Another source of qualitative data was the written assessments of the course, with suggestions for improvements, which students can submit at the end of the course in place of one missed homework assignment. Table 7.1 provides a summary of the changes planned for the course, and the source that provided motivation for the changes.

Table 7.1: Course improvement plan

Source	**Planned Changes**
MBKP and SPC application	• Improving Before and After questions • Changing lectures on o cash flow statement, o operating cash flow, o cash and book break-even
SEM model	• Reducing workload
SD model	• Reducing workload • Increasing lecture attendance
Student survey and end-of-the-course assessments	• Reducing number of homework assignments • Marking assignments on correctness and not on effort alone • Reducing number of problems on midterms • Reducing difficulty of problems on midterms • Including a number of worked-out problems at the end of each chapter of course notes • Eliminating take-home projects

7.3.2. Do

The changes were introduced in the undergraduate engineering management course taught during the Winter 2005 semester. Two sections of the course were available and were coded for this research as sections X and Y. Section X had an enrollment limit of 200 students, and section Y had a limit of 100 students. Lectures in both sections were delivered three times a week in the daytime.

While some changes differed between the two sections, the common changes included

- Reducing the number of homework assignments from 18 to 12. While the previous assignment schedule was irregular, in Winter 2005 there was one assignment per week, due on each Wednesday;
- Marking assignments on correctness instead of on effort alone. Previously, a student would receive a full mark even if a problem had not been solved correctly;
- Eliminating take-home projects. Previously, students had three take-home projects per semester, each worth 8% of the course grade. Students were required to work on projects without providing or receiving help, but during some semesters, several students were caught cheating and subjected to administrative action. Anecdotal evidence also suggested that cheating on the take-home projects was widespread. One student wrote in the end-of-the-course assessment:

  > "I think I was the only one who actually worked on the projects without help."

- Adding old homework problems as worked-out examples at the end of each chapter. Since the homework assignments were to be graded on correctness, the majority of the assigned problems were changed.

The changes specific to section Y included the re-design of the first midterm by of reducing the number of short-answer problems from 11 to 2, while the leaving number of true-false and multiple-choice problems the same.

In section X, the instructor monitored the knowledge transfer by using MBKPs with re-designed "Before and After" questions. The new questions were of approximately the same level of complexity among themselves, the terminology used in the questions was consistent with the terminology used in the course notes and during the lectures, and the question' structure was streamlined (for example, double negation was eliminated). The B&A questions appearing in the MBKPs administered during the Winter 2005 semester in Section X are presented in Appendix XVII. Instead of four B&A question per MBKP, the new MBKPs contained three B&A questions each.

Analysis of the SEM model presented in Chapter 5 suggested that the question "Do you own a personal computer" did not provide an adequate measure of a student's financial background. A student's gender originally was believed to provide no causal effects. While data suggested possible causal effects from gender on satisfaction with the course and the instructor (Section 5.6.2), these effects were attributed to sampling fluctuation and not to the true underlying causal relationships. Therefore, the survey described in Chapter 5 and presented in Table 5.3 was modified by excluding the questions "Do you own a personal computer?" and "What is your gender?"

7.3.3. Check

7.3.3.1. After-midterm survey

An after-midterm survey similar to one described in Chapter 5 (with the questions about gender and owning a computer deleted) was administered in two sections (X and Y) of the undergraduate engineering management course taught at the University of Alberta in the Winter 2005 semester. The response rates (as a percentage of the total number of students enrolled) were 64.6% in X, and 49% in Y. The overall combined sample size, after list-wise deletion of questionnaires with missing responses, was 165.

A number of statistical tests were carried out on the survey data. The details of these tests are presented in Appendix XVIII.

Since the students who participated in the survey were not likely to miss many lectures in the course (96% of respondents in section X and 98% of respondents in section Y reported that they missed "none" or "a few" lectures), it was decided to investigate whether the class sample participating in the survey was representative of the whole class.

It was decided to test whether the true midterm score distribution (available to the instructor from the midterm exam) corresponded to the distribution of midterm scores as reported by students on a survey. The frequency histograms (see Figures 7.1 and 7.2) indicated that the distribution of the reported midterm scores did not follow the distribution of the true midterm scores exactly.

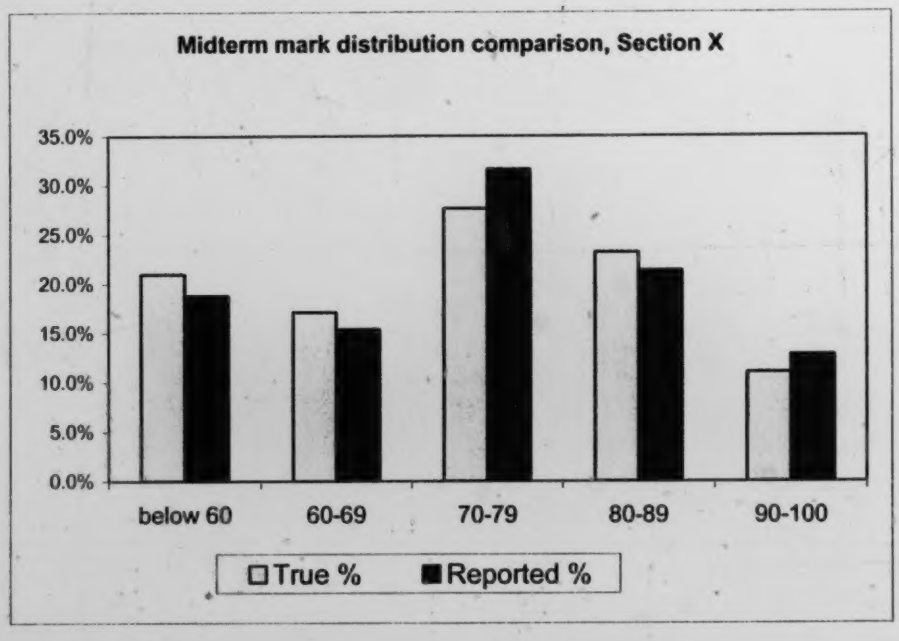


Figure 7.1: Midterm marks distribution, true and reported, section X

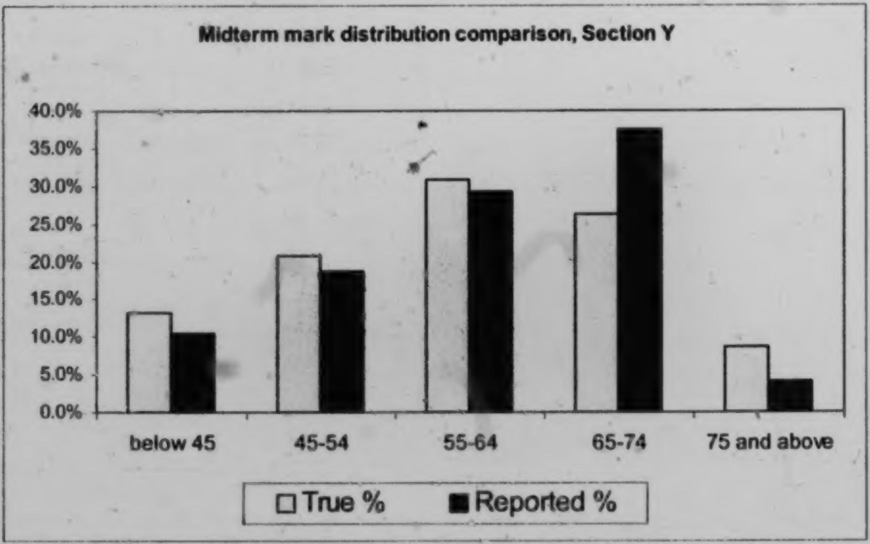


Figure 7.2: Midterm marks distribution, true and reported, section Y

A formal goodness-of-fit test (Montgomery and Runger 1999) was carried out to determine whether the distributions of the true and reported midterm marks were statistically different. The test results indicated no evidence that the reported and true midterm scores in either section came from different distributions ($P = 0.79$ for section X, $P = 0.43$ for section Y)

An interesting finding was that a higher proportion of students reported that it is important to have fun while being at university (99.48%) than that it is important to do well in university (95.38%). Similar proportions were observed in Fall 2003 – Winter 2004 (98.70% believed it is important to have fun, and 94.53% believed it is important to do well).

7.3.3.1.1. Comparison between Winter 2005 and Fall 2003 – Winter 2004

An analysis of the survey data indicated that the reduction in the number of assignments (in both sections) and in the midterm's difficulty (in section Y) reduced the perceived workload in the Winter 2005 sections of the course, as compared to the perceived workload of the previous years. The average perceived workload changed from 3.23 to 2.94, and the P value for the test was $P = 10^{-7}$. This change improved satisfaction with the instructor, and the increase was statistically significant at α=0.10 level. The average satisfaction with the instructor changed from 3.57 to 3.69 ($P = 0.089$). Satisfaction with the course did not change significantly (3.46 versus 3.33, $P = 0.105$), while the students' attitude toward the subject actually declined (from 3.66 to 3.36, $P = 0.0013$). The decline in attitude toward the subject highlights that other variables, besides satisfaction with the instructor and with the course, affect the students' attitude toward the subject. A more detailed investigation (in form of a SEM model) would be required to explain the changes in each variable.

7.3.3.1.2. Comparison between Sections X and Y, Winter 2005

Analysis of the difference in the perceived workload between sections X and Y indicated that while in section Y the workload was perceived as "average" or easier than average (100% of responses), in section X, 13% of the respondents rated the perceived workload as “high” or “very high.” Indeed, the perceived workload was significantly higher in section X than in section Y (an average perceived workload of 3.07 for section X versus 2.64 for section Y, $P = 10^{-6}$). The written comments pointed to a possible source of the perception that the workload was higher in section X: the quizzes, and the number of problems on the midterm exam:

> “It is also very unfair that [X] section has 4 (FOUR) (*sic*) quizzes in a semester. For me, the more tests I have, the more opportunity I have to lose marks”
>
> “Having 4 quizzes in addition to weekly assignments & midterms is a heavy workload”
>
> “Why do we need to be pressed for time during the exam?”
>
> “More time to write quizzes and test”

Moreover, as is often the case, the opposite views on each issue can always be found:

> “I like the weekly quizzes, they keep my studies up to date. I wish all my classes had quizzes like this!!!”

Despite the higher perceived workload in section X, the satisfaction with both the course and the instructor was higher in section X than in section Y. The average satisfaction with the course was 3.47 in section X and 3.00 in section Y ($P = 0.004$), and the average satisfaction with the instructor was 3.89 in section X and 3.21 in section Y ($P = 10^{-5}$). Instructor X drew some enthusiastic comments from the students:

> “One of the rare teachers that can make a course interesting!”
>
> “Instructor makes course interesting by using anecdotes and current events. This is good to spice things up a bit.”

While course organization received a higher value in section X, the students in section Y also commented on the organization of the course:

> "It is very helpful that the instructor supplies such a detailed course outline and notes. The notes can be used to reinforce the lectures and the assignments do this very well also since they flow with the lecture material so well. The best of this course is how extremely well organized it and the instructor are. No surprises."

While the attitude toward the subject was higher in section X (where the average attitude toward the subject was 3.41) than in section Y (where the average attitude toward the subject was 3.25), the difference was not statistically significant ($P = 0.176$).

7.3.3.1.3. Instructor Y, Winter 2004 and Winter 2005

The survey data indicated that the Section Y instructor performed worse than the Section X instructor (the students in Section Y reported lower satisfaction with the course and instructor). The lower performance ratings were caused, possibly, by Instructor Y's lesser teaching experience (2 years versus 7 years for instructor X). It was decided, then, to compare Instructor Y's performance relative, and not absolute, terms. Instructor Y taught a section of the same course in the Winter 2004 semester (instructor Y's Winter 2004 data became a part of the overall sample, which was analyzed by using an SEM model). To determine whether instructor Y's performance was improving in relative terms, instructor Y's data for the Winter 2004 semester were compared against the data for the Winter 2005 semester.

Statistical analysis showed that the perceived workload, a subject of numerous complaints, had become significantly lower (average perceived workload of 3.58 in the Winter 2004 and 2.64 in the Winter 2005, $P = 10^{-12}$). Satisfaction with the instructor had improved significantly (the average satisfaction with the instructor was 2.76 in the Winter 2004 and 3.21 in the Winter 2005, $P = 0.016$), while the increase in satisfaction with the course is statistically significant at $\alpha = 0.10$ level (average satisfaction with the course of 2.71 was the Winter 2004, and 3.00 in the Winter 2005, $P = 0.089$). The attitude toward the subject did not change significantly (3.30 versus 3.25, $P = 0.406$).

Using the analysis presented above, we can speculate that instructor Y is on a learning curve, and that, with time, instructor Y should improve.

7.3.3.2. SEM model

The validity of a statistical model may be questionable when the same sample is used both to estimate the model's parameters and to validate it. Some researchers argue that the best way to validate a statistical model is by testing it by using a second independent sample (Cooil et al. 1987).

It was decided, therefore, to test the SEM model constructed based on the Fall 2003 – Winter 2004 data (model "Final"), by using the survey data collected in the Winter 2005 semester. The model was modified by eliminating the concepts ξ_1 "Financial background", ξ_2 "Gender", and ξ_7 "Enrollment level" (since all students in sections X and Y were undergraduates), and all causal effects originating from ξ_7. The rest of the model's specifications remained the same as those of the SEM model "Final". The model, called "Winter 05", along with the table describing model's variables and their labels, are presented in a path-diagram form in Appendix XIX. The matrix of covariances between observed indicators (matrix S) is presented in Appendix XX, and syntax of the SEM model based on the Winter 2005 data is presented in Appendix XXI.

The model did not provide a satisfactory fit for the data ($\chi^2 = 143.09$, with 81 d.f., $P = 0.000$). Analysis of the model's output suggested that the addition of several causal effects might significantly improve the model fit (based on modification indexes). Some of those effects, in retrospect, should have been included in the model specifications. For example, the causal effect from the variable "Importance of having fun while in university" on the variable "Time devoted to self-studying" might significantly improve the model fit (modification index of 20.9). The causal effect would take on a value of

–2.47, indicating that if a student believes that having time for fun while studying at a university is important, this student will spend less time studying.

A modification index of 5.08 for a causal effect of the variable "Extra-curricular activities" on the concept "Satisfaction with the course," and effect's expected value of 0.20, suggested that if a student participated in more extracurricular activities, this student's satisfaction with the course would improve. The causal relationship appears to be reasonable if we assume that increased participation in extra-curricular activities is possible if a student has more free time, which a student will have if the course workload is light. It was expected that this relationship would work through the "Extra-curricular activities" – "Time devoted to self-studying" – "Perceived workload" – "Satisfaction with the course" chain of causal effects. However, the causal effects of "Time devoted to self-studying" on "Perceived workload," and of "Perceived workload" on "Satisfaction with the course" were statistically insignificant. The failure of the causal chain required a direct effect of "Extra-curricular activities" on "Satisfaction with the course." A number of modification indexes pointed to the conclusion that the coordination between the concepts "Extra-curricular activities" and "Satisfaction with the instructor" is also not well explained by the model, possibly due to the same reasons as for the coordination between the concepts "Extra-curricular activities" and "Satisfaction with the course" (a causal chain with non-significant effects).

Model diagnostics also suggested a strong causal effect of the variable "Instructor's experience" on the variable "Perceived workload" (modification index of 15.7). This effect might have been a data artifact and not a true causal relationship. As well, a more experienced instructor might better understand what level of workload (determined by the assignment rate, assignment difficulty, midterm difficulty, number of quizzes, etc.) is appropriate for students. It was shown in the previous section that while the perceived workload was higher in section X than in section Y, satisfaction with the instructor and with the course was also higher. The effect of the instructor's experience on the perceived workload is, possibly, transmitted via a more complex causal chain, but for this model, it was decided that instructor's experience indeed played a role in determining the proper level of the workload for the course (and, therefore, the students' perceived workload).

Therefore, the four causal effects were added to the model in an attempt to improve the model's fit: from "Importance of having fun while in university" to "Time devoted to self-studying," from "Extra-curricular activities" to "Satisfaction with the course," from "Extra-curricular activities" to "Satisfaction with the instructor," and from "Instructor's experience" to "Perceived workload." The modified model "Winter 05" fit the data appropriately ($\chi^2 = 98.28$, with 77 d.f., $P = 0.051$). Analysis of the model's output did not show any serious problems with the model. The largest modification index for the B matrix was 4.16, and 4.49 for the Γ matrix. The residuals were distributed normally, along the 45° line on the normal probability plot. The largest absolute value of a standardized residual was 2.42.

7.3.3.3. Knowledge transfer

The process of knowledge transfer from the instructor to students during lectures was monitored in section X of the Winter 2005 course, using the Modified Background Knowledge tool and a Statistical Control Chart to analyze the process. The data collected from the MBKPs are presented in Appendix XXII.

By using the approach described in Chapter 4, a SPC chart displaying the statistics "Proportion Incorrect Before, Correct After" P_{IBCA} and "Incorrect After" P_{IA} was constructed (see Figure 7.3).

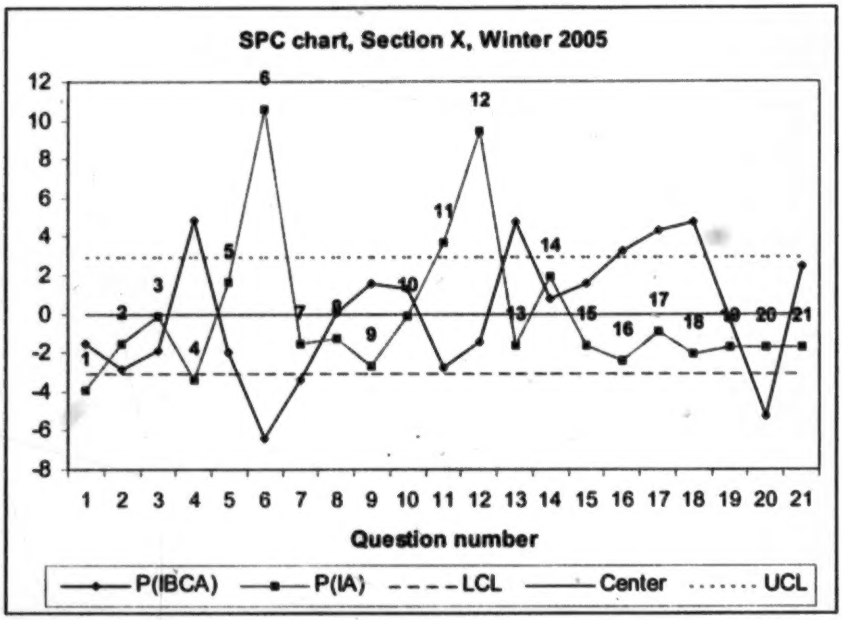


Figure 7.3: Knowledge transfer SPC chart, Instructor X, Winter 2005

The process of knowledge transfer was, again, not in statistical control. For example, for question 6, P_{IA} plotted far above the upper control limit (UCL), and P_{IBCA} plotted below the lower control limit (LCL), suggesting that low knowledge gain had occurred for the subject matter corresponding to the question. Analysis of the question itself, and retrospective analysis of the lecture, suggested that two factors might have contributed to the out-of-control situation. Firstly, the material corresponding to the question was not covered in sufficient detail during the lecture, and, secondly, the question (by mistake) had two correct answers. Question 4, on the other hand, indicated that a high number of students had learned the subject matter covered by the question (P_{IBCA} above UCL, and P_{IA} below LCL). Analysis of the question and lecture suggested that the subjects (decision trees and expected monetary outcome) had not been new to the students, and that they possibly had had background knowledge of the material.

While the process of knowledge transfer was not in statistical control, a significant improvement in knowledge gain occurred when the data were compared to those of the 2002-2003 semesters. In the Winter 2002 semester, instructor A (the same instructor teaching section X in Winter 2005) produced the best knowledge transfer results ($P_{IBCA} = 0.276$, $P_{IA} = 0.189$). These results were compared with those of the Winter 2005 performance. (Table 7.2):

Table 7.2: Absolute difference in performance, 2002-2003 and 2005

Statistic	Winter 2005	Winter 2002	Absolute difference	Weighted *p*	Z	Statistically significant?
P_{IBCA}	0.555	0.276	0.279	0.426	16.59	Yes
P_{IA}	0.163	0.189	-0.027	0.175	-2.06	Yes

The results indicate that the knowledge gained increased (see the significant increase in P_{IBCA}), while, at the same time, the proportion of students not benefiting from the lecture decreased (see the significant decrease in P_{IA}).

7.3.4. Act

The changes introduced before and during the Winter 2005 semester of teaching the course produced some significant results. The perceived course workload became lower, and the overall satisfaction with the instructor improved during Winter 2005, compared to the years 2003-2004.

When performance within the sections of the course is analyzed against that of the previous years, the changes are even more profound. For instructor Y, satisfaction with the instructor and with the course significantly improved, compared to that of the previous year. In section X, the changes to the B&A questions and the MBKP administration significantly improved the process of knowledge transfer.

A structural model constructed based on the 2003-2004 data proved to be robust enough, despite the changes in the underlying student population. With several minor changes, the model provided a satisfactory fit to the Winter 2005 data. While estimates of some structural coefficients changed, the overall model structure still adequately described the classroom educational system.

Analysis of the written comments on the Winter 2005 survey provided new ideas for course-improvement opportunities. The midterm exam and quiz difficulty can be adjusted to give students enough time to demonstrate their knowledge, without sacrificing test integrity. Further changes have been suggested for the course organization:

- Improve coordination between course notes and course classroom presentations
- Provide a glossary of terms and their synonyms
- Better organization of the course notes
- Better organization of the course syllabus
- Regular instructor's office hours

One of the goals of the introduced changes was to improve the students' attitude toward the subject. According to the SEM model, a lower perceived workload was expected to increase satisfaction with the instructor (a results which did occur) and satisfaction with the course (a result which did not occur). An increase in satisfaction, in turn, was expected to improve the attitude toward the subject. As was illustrated earlier, this result did not happen.

Figure 6.24 illustrates the dynamic simulation behavior of the "Attitude toward the subject" variable under different assignment workload scenarios. The workload in 2003-2004 corresponded to scenario 3 (1.5 assignments per week, 3 hours of homework per assignment), while during Winter 2005, the workload corresponded to scenario 2 (1 assignment per week, 3 hours of homework per assignment). The simulation indicated that half-way through the course when the survey was administered, the attitude toward the subject should have been higher under scenario 2.

The SD model was based on LISREL estimates of causal effects. The SEM model based on 2003-2004 data explained 55.3% of the variance in the concept "Attitude toward the subject," and the SEM model based on the Winter 2005 data explained 62.7% of the variance. Forces unaccounted for by the SEM models might have produced a lower attitude toward the subject despite the significantly lower perceived workload.

One such force could have been a negative predisposition against the course, formed by the students even before they took the course. Engineering undergraduate students at the University of Alberta are required to take a course in engineering management in order to graduate, and, therefore, some of the students may be "unwilling participants" in the course. Some written comments indicated that the students might have been projecting their negative predisposition against the course onto the subject matter:

> **"I have many other courses that have a higher impact on my career and I spend my time on those courses instead"**
>
> **"I probably will never use this knowledge in my future career"**

Since for many students this course may be their first and last exposure to the field of finance, accounting, and engineering management (at least during their university years), students must form a positive impression of both the course and the instructor from the very beginning. The instructor simply may not have enough time to change that impression.

7.4. Afterword

While principles of continuous improvement require continuous monitoring of customer satisfaction, one must not lose the sight of the primary function of the instructor in a classroom – teaching. Student comments suggest that too much effort devoted to measuring and monitoring classroom performance is just as bad as not enough effort:

> "This kind of useless survey takes away from lectures, it's all pretty useless."
>
> "Too many surveys, questionnaires & analysis of the course. Spend more time actually teaching".

7.5. Summary

Chapter 7 described a systematic continuous approach to improving the quality of an undergraduate engineering management course. The quality improvement plan was formulated based on insight gained from the models presented in Chapters 4, 5, and 6, and from the qualitative feedback provided by students. The changes were implemented during the Winter 2005 semester and evaluated by using the same tools – MBKPs and SPC, SEM, and student surveys. Analysis of the Winter 2005 data indicated that while some objectives had been achieved (i.e., reducing the perceived workload, improving satisfaction with the instructor), some had not (i.e., improving students' satisfaction with the course, improving their attitude toward the subject).

Since quality improvement is an iterative process, the next step will be to start again along the path laid out in Figure 3.1, starting with the identification of new input, process, and outcome candidates for performance measurement.

Chapter 8: Conclusions

8.1. Contributions of the research

Chapter 3 introduced an educational performance framework, based on the systems approach, for measuring educational performance at the classroom level. This performance framework covers the three most important elements of an educational system at the classroom level: the course, students, and instructor. This framework is balanced yet comprehensive with a number of performance indicators for each element of the system. The performance framework was created while keeping in mind the problems and challenges of measuring performance in the non-profit sector. The number of performance indicators within each aspect of the educational performance framework can be increased or decreased to accommodate individual instructor's preferences. This performance framework integrated such quality management tools as the Statistical Process Control, the Structural Equation Modeling, and the Systems Dynamics in order to bring performance measurement into a post-secondary classroom.

In Chapter 4, the process of knowledge transfer between an instructor and students was identified as a candidate for continuous monitoring. A tool for collecting data on knowledge transfer, called the "Modified Background Knowledge Probe," was introduced. The methodology for creating, administering, and analyzing the MBKP data was described, including suggestions for the Before and After questions design, methods of resolving the problems of autocorrelation, short run, and warm-up instability, and approaches for searching for the assignable causes of the out-of-control situations. The problem of small class size (producing a small sample size) was investigated by introducing a variable knowledge transfer statistic.

In Chapter 5, a methodology called "Structural Equation Modeling" was used for discovering cause-and-effect relationships in a classroom. A model based on hypotheses about the relationships among the course, student, and instructor characteristics confirmed the hypotheses of the presence of the feedback loops in the classroom system. The model allowed for accounting for the imprecise nature of measuring "soft" variables such as attitude and satisfaction, and for the limitations of data obtained through observational study.

In Chapter 6, a dynamic hypothesis about classroom system's behaviour was postulated based on the insights gained in Chapter 5 on the relationships among the variables involved in the feedback loops. For the first time an SD model was created using the estimates of the relationships among the system's variables obtained from the SEM model. The System Dynamics modeling approach was used to examine the changes in system variables over time. The System Dynamics model was used to find "leverage points" – model elements where small changes in inputs lead to significant changes in outputs. Policies aimed at improving student performance and attitudes were tested by modifying the parameters that are under influence of an instructor: homework difficulty and lecture attendance.

Chapter 7 illustrated how a Plan-Do-Check-Act continuous improvement approach can be applied to managing educational performance at the classroom level. The understanding of the classroom system gained through the system modeling and analysis together with student feedback were used to design changes in an undergraduate engineering management course. Analysis of quantitative and qualitative data suggested that the introduced changes had improved some of the aspects of classroom performance. Opportunities for future improvement were identified as well.

8.2. Directions of future research

The work carried out and described in Chapters 4 to 7 completed the first iteration along the methodology path in Figure 3.1. Future research should be directed at refining the presented system of measuring, modeling, and managing an educational system at the classroom level.

The performance framework should be analyzed to determine whether it needs to be expanded to include additional instructor- and course-related indicators, or whether some indicators can be excluded to reduce the data collection and analysis workload.

Regarding the application of the statistical process control in education, other processes suitable for continuous monitoring at the classroom level must be identified. If such processes exist, tools for data collection and analysis must be designed. A typical progression in quality control implementation is to move from control charts for attributes to control charts for variables as knowledge about key processes and variables becomes available (Montgomery 1997a). Therefore, further research efforts should be concentrated on developing variable process statistics. Doing so will not only allow for better quality control, but will also help to address the problems of small class size and small number of data samples. Further research is also necessary in the use of designed experiments in finding the assignable causes of poor knowledge transfer, and the reason behind the out-of-control situation when it is not obvious.

The presented structural equation model of an educational classroom system was estimated with an implicit assumption that the mean values of the model's indicators and concepts are equal to zero. This condition was achieved by using a matrix of covariances among the observed indicators – a matrix that records deviations of indicators from their means. The estimation of SEM models without means stems, partially, from the path analysis tradition, and, partially, from additional data and formulation requirements that a model with means would introduce (Hayduk 1987). Using the output of a "no means" model, we can estimate the expected changes in the dependent variables given a specified change in the independent variables (e.g., $Var(Y) = b^2 Var(X)$), but we cannot predict what the expected value of a dependent variable $E(Y)$ will be (i.e., we cannot estimate $E(Y) = a + b*E(X)$). Future research efforts should be focused on introducing means into the SEM model. Doing so will provide several benefits: better test of the overall model, better estimates of structural coefficients, and the prediction of the values of the dependent variables for any combination of the values of the independent variables (Hayduk 1987).

Future research in the application of system dynamics at the classroom level should concentrate on expanding the model's boundary. A future model should include, in its feedback structure, variables that are currently assumed to remain constant throughout the model's time horizon.. For example, the expected test score might change depending on the level of workload and satisfaction. Additionally, presently it is assumed that the effects of the structural causal coefficient will remain linear regardless of the value of the input and output variables. Research into the non-linearity of causal effects is necessary. The expansion of the model's boundary should include feedback structures involving the instructor and the course structure. For example, assignment difficulty may depend on the students' current level of performance and satisfaction, while instructional quality may depend on the instructor's own satisfaction with the students' performance.

Additional effort is needed to discover and model the factors contributing to the students' attitude toward the subject. The variable "Attitude toward the subject" had one of the highest proportions of explained variance in the SEM models, but it was, possibly, the variables that were not included in the model that affected the attitude toward the subject during the Winter 2005 semester.

The SD model was created by using quantitative data. Researchers are becoming increasingly aware of the importance of collecting qualitative data on the quality of educational processes (Jones 2003), and of incorporating qualitative data into the SD models (Luna-Reyes and Andersen 2003, Coyle 2000). Additional research is needed on modeling educational systems by using qualitative data.

Another area for investigation could include the introduction of a quantitative test to assess the quality of a system dynamics model. The test can be modeled on a chi-square test used in structural equation modeling. An SD model produces a particular set of output values given a particular set of input values, in effect reproducing the behaviour of a single individual (or group of individuals with similar characteristics). By running multiple sets of initial values in the SD model, we can obtain a matrix of variances and covariances between the model's input and output variables. This matrix would constitute a model-implied set of variances and covariances, similar to SEM's matrix Σ. This matrix can be compared to the matrix of observed variances and covariances among the model's variables (similar to SEM's matrix *S*). The two matrices can be compared by using a chi-squared test to evaluate how well the system dynamics model can recreate data coming from the real world. Even though replication of a historical behavior is no guarantee that the model will be able to correctly predict the future (Jackson 2000, Legasto and Maciariello 1980), a historical behavior replication test is routinely included in model testing (Sterman 2000, Forrester and Senge 1980).

Bibliography

1. Abbott, R.D., Wulff, D.H., Nyquist, J.D., Ropp, V.A., and Hess, C.W., 1990. "Satisfaction With Process of Collecting Student Opinions About Instruction: The Student Perspective", *Journal of Educational Psychology,* Vol. 82, No. 2, pp. 201 – 206.

2. AISI, 1999. *Framework for the Alberta Initiative for School Improvement.* Alberta Initiative for School Improvement Education Partners Steering Committee, Alberta Learning, Canada. Available on-line at: http://www.learning.gov.ab.ca/sib/aisi

3. Alberta Learning, 2001. *Results Report on Alberta's Learning System, 2000/2001.* Alberta Learning, Government of Alberta, Canada.

4. Alberta Learning, 2002. *Guide for School Board Planning and Results Reporting,* Alberta Learning, Government of Alberta, Canada

5. Alberta Finance, 1996. *Measuring Performance—A Reference Guide,* Alberta Finance, Government of Alberta, Canada. Available on-line at: http://www.finance.gov.ab.ca/publications/measuring/index.html

6. Alberta Finance, 2001. *Measuring Up: 2000-2001 Annual Report*, Alberta Finance, Government of Alberta, Canada. Available on-line at: http://www.finance.gov.ab.ca/publications/measuring/index.html

7. Andersen, B., and Fagerhaug, T., 2002. *Performance Measurement Explained: Designing and Implementing Your State-of-the Art System*, ASQ Quality Press, Milwaukee, WI, USA.

8. Andersen, D.F., and Richardson, G.P., 1980. "Toward Pedagogy of System Dynamics", appears in Legasto, A.A., Forrester, J.W., and Lyneis, J.M., editors, 1980. *System Dynamics.* Studies in the Management Sciences, Vol.14, North-Holland Publishing Company, New York, NY, USA

9. Angelo, T.A., and Cross, K.P., 1993. *Classroom Assessment Techniques: A Handbook for College Teachers*; Jossey-Bass Publishers: San-Francisco, CA, USA.

10. Arias, J.J., and Walker, D.M., 2004. "Additional Evidence on the Relationship Between Class Size and Student Performance", *Journal of Economic Education,* Vol. 35, No. 4, pp. 311 – 330.

11. Arquitt, S., and Johnstone, R., 2004. "A scooping and consensus building model of a toxic blue-green algae bloom", *System Dynamics Review,* Vol. 20, No. 2, pp. 179-198.

12. Baker, B.D., and Richards, C.E., 2002. "Exploratory Application of System Dynamics Modeling To School Policy Analysis", *Journal of Educational Finance,* Vol. 27, No. 1, pp. 857-883.

13. Bartholomew, D.J., 1983. "Latent Variable Models For Ordered Categorical Data", *Journal of Econometrics,* Vol. 22, pg. 229-243.

14. Bell, J.A., and Senge, P.M., 1980. "Methods for enhancing refutability in System Dynamics modeling", appears in Legasto, A.A., Forrester, J.W., and Lyneis, J.M., editors, 1980. *System Dynamics.* Studies in the Management Sciences, Vol.14, North-Holland Publishing Company, New York, NY, USA.

15. Berman, E., and Wang, X., 2000. "Performance measurement in US counties: Capacity for reform", *Public Administration Review,* Vol.60 No.5, pp.409-420.

16. Besterfield-Sacre, M., Amaya, N.Y., Shuman, L.J., and Atman, C.J., 1998. "Implications of Statistical Process Monitoring for ABET 2000 Program Evaluation: An Example Using Freshman Engineering Attitudes", *American Society for Engineering Education Conference Proceedings*, Seattle, WA, June 1998, (CD ROM).

17. Bititci, U. S., Carrie, A. S., and McDevitt, L., 1997. "Integrated performance measurement systems: a development guide", *International Journal of Operations and Production Management,* Vol. 17, No.5, pp.522-534.

18. Bjerke, F., Aastveit, A.H., Stroup, W.W., Kirkhus, B., and Næs, T., 2004. "Design and Analysis of Storing Experiments: A Case Study", *Quality Engineering*, Vol. 16, No. 4, pp. 591-611.

19. Blackmur, R., 2004. "Issues in higher education quality assurance", *Australian Journal of Public Administration,* Vol. 63, No. 2, pp. 105-116.

20. Boland, T., and Fowler, A., 2000. "A systems perspective of performance management in public sector organizations", *International Journal of Public Sector Management,* Vol. 13, No. 5, pp. 417-446.

21. Bourne, M., Mills, J., Wilcox, M., Neely, A., and Platts, K., 2000. "Designing, implementing and updating performance measurement systems", *International Journal of Operations & Production Management,* Vol.20, No.7, pp. 754-771.

22. Browne, M.W., and Cudeck, R., 1992. "Alternative Ways of Assessing Model Fit", *Sociological Methods and Research,* Vol. 21, No. 2, pp. 230-258.

23. Burke, J. C., 2003. "Trends in higher education performance", *Spectrum,* Vol. 76, No. 2, pp. 23-24.

24. Cairney, R., 2002. "The cost of excellence", *Folio,* University of Alberta, Vol. 40, No. 7, p.3.

25. CAUT, 2005. *Almanac of Post-Secondary Education in Canada,* Canadian Association of University Teachers.

26. Cavana, R.Y., and Mares, E.D., 2004. "Integrating critical thinking and systems thinking: from premises to causal loops", *System Dynamics Review,* Vol. 20, No. 3, pp. 223 – 235.

27. Chang, O.H., and Chow, C.W., 1999. "The Balanced Scorecard: A potential Tool for Supporting Change and Continuous Improvement in Accounting Education", *Issues in Accounting Education,* Vol. 14, No. 3, pp. 395-412.

28. Cochran, D.J., and Riley, M.W., 1986. "An Evaluation of Knife Handle Guarding", *Human Factors*, Vol. 28, Vol. 3, pp. 295-301.

29. Conley, D.T., and Picus, L.O., 2003. "Oregon's Quality Education Model: Linking Adequacy and Outcomes", *Educational Policy,* Vol. 17, No. 5, pp. 586-612.

30. Cooil, B., Winer, R.R., and Rados, D.L., 1987. "Cross-Validation for Prediction", *Journal of Marketing Research,* Vol. 24, No. 3, pp. 271-279.

31. Coyle, G., 2000. "Qualitative and quantitative modeling in system dynamics: some research questions", *System Dynamics Review,* Vol. 16, No. 3, pp. 225-244.

32. Cranton, P., and Smith, R.A., 1990. "Reconsidering the Unit of Analysis: A Model of Student Ratings of Instruction", *Journal of Educational Psychology,* Vol. 82, No. 2, pp. 207-212.

33. Cullen, J., Joyce, J., Hassall, T., and Broadbent, M., 2003. "Quality in Higher Education: From Monitoring to Management", *Quality Assurance in Education,* Vol. 11, No. 1, pp. 5-14.

34. Cunningham, G.B., Sagas, M., Dixon, M., Kent, A., and Turner, B.A., 2005. "Anticipated Career Satisfaction, Affective Occupational Commitment, and Intentions to Enter the Sport Management Profession", *Journal of Sport Management,* Vol. 19, No. 1, pp. 43-57.

35. De Lancer Julnes, P., and Holzer, M., 2001. "Promoting the Utilization of Performance Measures in Public Organizations: An Empirical Study of Factors Affecting Adoption and Implementation", *Public Administration Review,* Vol.61, No.6, pp.693-708.

36. Deming, W.E., 1986. *Out of the Crisis,* Massachusetts Institute of Technology, Center for Advanced Engineering Study, Cambridge, MA, USA.

37. Dudley, R.G., 2004. "Modeling the effects of a log export ban in Indonesia", *System Dynamics Review,* Vol. 20, No. 2, pp. 99-116.

38. Duncan, O.D., 1975. *Introduction to Structural Equation Models,* Academic Press, Burlington, MA, USA.

39. Eftekhar, N., 1998. *Dynamic Modeling of a Teaching/Learning System To Aid System Re-engineering,* PhD thesis, University of Manitoba, Winnipeg, Manitoba, Canada.

40. Elmore, R.F., and Rothman, R., editors, 2000. *Testing, Teaching, and Learning. A Guide for States and School Districts,* Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education, National Research Council, National Academy Press, Washington, DC, USA.

41. Fitz-Gibbon, C. & Kochan S., 2000. School Effectiveness and Education Indicators. In Teddlie, C. & Reynolds, D., 2000. *The International Handbook of School Effectiveness Research,* Falmer Press, New York, USA.

42. Flower, D., 2004. "Public Education as Trojan Horse: the Alberta Case", *Alberta Teacher's Association Magazine,* Vol. 84, No. 4, pp. 4-12.

43. Foltin, C., 1999. "State and local government performance: It's time to measure up!", *The Government Accountants Journal,* Vol.48 No.1, pp.40-46.

44. Forrester, J.W., 1968. *Principles of Systems,* 2nd preliminary edition, Wright-Allen Press, Inc., Cambridge, MA, USA.

45. Forrester, J.W., 1980. "System dynamics—future opportunities", appears in Legasto, A.A., Forrester, J.W., and Lyneis, J.M., editors, 1980. *System Dynamics.* Studies in the Management Sciences, Vol.14, North-Holland Publishing Company, New York, NY, USA.

46. Forrester, J.W., 1989. "The Beginning of System Dynamics", Banquet Talk at the international meeting of the System Dynamics Society, Stuttgart, Germany, July 13, 1989. Available from http://sysdyn.clexchange.org/sdep/papers/D-4165-1.pdf

47. Forrester, J.W., and Senge, P.M., 1980. "Building confidence in system dynamics models", appears in Legasto, A.A., Forrester, J.W., and Lyneis, J.M., editors, 1980. *System Dynamics.* Studies in the Management Sciences, Vol.14, North-Holland Publishing Company, New York, NY, USA.

48. Fortenberry, N.L., 1999. "An Educational Research Agenda for SMET Higher Education", *American Society for Engineering Education Conference Proceedings,* Charlotte, NC, USA.

49. GAO, 1980. U.S. General Accounting Office, *Evaluating a performance measurement system: a guide for the Congress and Federal agencies: report to the chairwoman, Subcommittee on Civil Service, House Committee on Post Office and Civil Service / by the Comptroller General of the United States,* (GAO/FGMSD-80-57), Washington, D.C., U.S. General Accounting Office.

50. George, R.and Kaplan, D., 1998. "A Structural Model of Parent and Teacher Influences on Science Attitudes of Eight Graders: Evidence from NELS: 88", *Science Education*, Vol. 82, pp. 93-109.

51. Ghobadian, A., and Ashworth, J., 1994. "Performance measurement in local government—concept and practice", *International Journal of Operations & Production Management,* Vol.14 No.5, pp.35-51.

52. Glover, M., 1992. *A Practical Guide for Measuring Program Efficiency and Effectiveness in Local Government,* An Innovation Groups Publication, Tampa, FL.

53. Griffin, B.W., 2004. "Grading leniency, grade discrepancy, and student rating of instruction", *Contemporary Educational Psychology,* Vol. 29, No. 4, pp. 410-425.

54. Griffin, P., Coates, H., McInnis, C., and James, R., 2003. "The Development of an Extended Course Experience Questionnaire", *Quality in higher Education,* Vol. 9, No. 3, pp. 259-266.

55. Grizzle, G.A., and Pettijohn, C.D., 2002. "Implementing Performance-Based Program Budgeting: A System-Dynamics Perspective", *Public Administration Review,* Vol. 62, No. 1, pp. 51-62.

56. Grygoryev, K., and Karapetrovic, S., 2005. "An integrated system for educational performance measurement, modeling, and management at the classroom level", *Total Quality Management Magazine,* Vol. 17, No. 2, pp. 121-136.

57. Hahs-Vaughn, D., 2004. "The impact of Parents' Education Level on College Students: An Analysis Using the Beginning Post-secondary Students Longitudinal Study 1990-92/94", *Journal of College Student Development,* Vol. 45, No. 5, pp. 483-500.

58. Harris, A., 1998. "Effective Teaching: a Review of the Literature", *School Leadership & Management,* Vol. 18, No. 2, pp. 169 – 183.

59. Harris, H., and Bretag, T., 2003. "Reflective and Collaborative Teaching Practice: working towards quality student learning outcomes", *Quality in Higher Education,* Vol. 9, No. 2, pp. 179-185.

60. Harvey, L., 2003. "Student Feedback", *Quality in Higher Education,* Vol. 9, No. 1, pp. 3-20.

61. Hatry, H.P., 1999. *Performance Measurement: Getting Results,* The Urban Institute Press, Washington, D.C.

62. Haveman, R., and Wolfe, B., 1995. "The Determinants of Children's Attainment: A Review of Methods and Findings", *Journal of Economic Literature*, Vol. 33, No. 4, pp. 1829 – 1878.

63. Hayduk, L.A., 1987. *Structural Equation Modeling with LISREL, Essentials and Advances.* The John Hopkins University Press, Baltimore, MD, USA.

64. Hayduk, L.A., 1994. "Personal Space: Understanding the SIMPLEX Model", *Journal of Nonverbal Behavior*, Vol. 18, No. 3, pp. 245-260.

65. Hayduk, L.A., 1996. *LISREL Issues, Debates, and Strategies.* The John Hopkins University Press, Baltimore, MD, USA.

66. Hayduk, L.A., Stratkotter, R.F., and Rovers, M.W., 1997. "Sexual Orientation and the Willingness of Catholic Seminary Students to Conform to Church Teachings", *Journal for the Scientific Study of Religion,* Vol. 36, No. 3, pp. 455-467.

67. Hayduk, L., Cummings, G., Stratkotter, R., Nimmo, M., Grygoryev, K., Dosman, D., Gillespie, M., Pazderka-Robinson, H., and Boadu, K., 2003. "Pearl's D-Separation: One More Step Into Causal Thinking", *Structural Equation Modeling*, Vol. 10, No. 2, pp. 289-311.

68. Hayduk, L.A., Pazderka-Robinson, H., Cummings, G., Levers, M.D., and Beres, M.A., 2005. "Structural Equation Model Testing and the Quality of Natural Killer Cell Activity Measurements", *BMC Medical Research Methodology,* Vol. 5, No. 1, available from http://www.biomedcentral.com/1471-2288/5/1.

69. Heise, D.R., 1975. *Causal Analysis,* Wiley-Interscience Publication, John Wiley & Sons, New York, NY, USA.

70. Hillier, F.S., 1969. "X- and R-Chart Control Limits Based on A Small Number of Subgroups", *Journal of Quality Technology*, Vol.1 No.1, pp.17-25

71. Hillison, W.A., Hollander, A.S., Icerman, R.C, and Welch, J. 1995. *Use and audit of performance measures in the public sector.* The Institute of Internal Auditors Research Foundation, Florida, USA.

72. Homer, J., Hirsch, G., Minniti, M., and Pierson, M., 2004. "Models for collaboration: how system dynamics helped a community organize cost-effective care for chronic illness", *System Dynamics Review,* Vol. 20, No. 3, pp. 199 – 222.

73. Hox, J.J., and Bechger, T.M., 1998. "An Introduction to Structural Equation Modeling", *Family Science Review,* Vol. 11, pp. 354-373.

74. Isely, P., and Singh, H., 2005. "Do Higher Grades Lead to Favorable Student Evaluations?", *Journal of Economic Education*, Vol. 36, No. 1, pp. 29-42.

75. ISO 9000 (2000). *Quality management systems – Fundamentals and vocabulary,* International Organization for Standardization, Geneva, Switzerland.

76. ISO 9001 (2000), *International Standard: Quality Management Systems: Requirements,* International Organization for Standardization, Geneva, Switzerland.

77. Jackson, M.C., 2000. *Systems Approaches to Management,* Kluwer Academic/Plenum Publishers, New York, NY, USA.

78. Jenkins, Lee, 2003. *Improving Student Learning: Applying Deming's Quality Principles In Classrooms,* 2nd ed., ASQ Quality Press, Milwaukee, WI, USA.

79. Jones, S., 2003. "Measuring the Quality of Higher Education: linking teaching quality measures at the delivery level to administrative measures at the university level", *Quality in Higher Education,* Vol. 9, No. 3, pp. 223-229.

80. Jöreskog, K., and Sörbom, D., 1996a. *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*, Scientific Software International, Chicago, IL, USA.

81. Jöreskog, K., and Sörbom, D., 1996b. *LISREL 8: User's Reference Guide*, Scientific Software International, Chicago, IL, USA.

82. Kaplan, R.S., 2001. "Strategic performance measurement and management in non-profit organizations", *Nonprofit Management and Leadership,* Vol.11 No.3, pp.353-370.

83. Kaplan, R.S., and Norton, D.P., 2001. *The Strategy-Focused Organization.* Harvard Business School Press, Boston, MA.

84. Karapetrovic, S., 1998. *Quality Assurance in The University System*, PhD thesis, University of Manitoba, Winnipeg, Manitoba, Canada.

85. Karapetrovic, S., 2002. "Why and How to Develop a Meaningful Quality Assurance System in Engineering Schools", *International Journal of Engineering Education*, Vol.18 No.3, pp. 285-294.

86. Karapetrovic, S. and Rajamani, D., 1998. "An Approach to the Application of Statistical Quality Control Techniques in Engineering Courses", *Journal of Engineering Education,* Vol.82 No.2, pp. 269-276.

87. Karapetrovic, S., and Willborn, W. 1998. "The system's view for clarification of quality vocabulary", *International Journal of Quality and Reliability Management,* Vol.15 No.1, pp.99-120.

88. Karapetrovic, S., Rajamani, D., and Willborn, W., 1998. "ISO 9001 Quality System: An Interpretation for the University", *International Journal of Engineering Education,* Vol. 14, No. 2, pp. 105-118.

89. Karbani, T., 2005. "English department proposes half-year course for first-year students", *The Gateway,* University of Alberta, Vol. XCIV, Issue 26, 13 January 2005.

90. Kelton, W.D., Sadowski, R.P., and Sadowski, D.A., 1998. *Simulation with Arena.* WCB McGraw-Hill, Boston, MA, USA.

91. Kent, T.W., and Hasbrouck, R.B., 2003. "The Structural Factors That Affect Classroom Team Performance", *Team Performance Management,* 2003, Vol. 9, No 7/8, pp. 161-166.

92. Kripalani, M., 2004. "Getting The Best To The Masses", *Business Week,* Issue 3903, October 11, 2004, pg. 174.

93. Lane, K.L., Bocian, K.M., MacMillan, D.L., and Gresham, F., M., 2004. "Treatment Integrity: An Essential But Often Forgotten Component of School-Based Interventions", *Preventing School Failure,* Vol. 48, No. 3, pp. 36-44.

94. Legasto, A.A., and Maciariello, J., 1980. "System Dynamics: A Critical Review", appears in Legasto, A.A., Forrester, J.W., and Lyneis, J.M., editors, 1980. *System Dynamics.* Studies in the Management Sciences, Vol.14, North-Holland Publishing Company, New York, NY, USA.

95. Levine, D.M., Berenson, M.L., and Stephan, D., 1998. *Statistics for Managers using Microsoft Excel,* Prentice Hall: Upper Saddle River, NJ, USA.

96. Luna-Reyes, L.P., and Andersen, D.L., 2003. "Collecting and analyzing qualitative data for system dynamics: methods and model", *System Dynamics Review*, Vol. 19, No. 4, pp. 271-296.

97. MacLean's, 2004. *MacLean's Universities Ranking 2004,* available on-line at http://www.macleans.ca/universities/

98. Marks, R.B., 2000. "Determinants of Student Evaluations of Global Measures of Instructor and Course Value", *Journal of Marketing Education,* Vol. 22, No. 2, pp. 108-119.

99. Marsh, H.W., 1987. "Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research", *International Journal of Educational Research,* Vol. 11, No.3, pp. 253-388.

100. Meijer, R.R., 2002. "Outliers Detection in High-Stakes Certification Testing", *Journal of Educational Measurement*, Vol. 39, No. 3, pp. 219 – 233.

101. McLellan, A., 2005. *The Anne McLellan Report – Winter 2005,* Member of Parliament report to constituents, Vol. 4, No. 2.

102. McKeachie, W.J., 1990. "Research on College Teaching: The Historical Background", *Journal of Educational Psychology,* Vol. 82, No. 2, pp. 189-200.

103. Meade, P., Morgan, M., and Heath, C., 1999. "Equipping Leaders to Capitalize on the Outcomes of Quality Assessment in Higher Education". *Assessment & Evaluation in Higher Education*, Vol. 24, No. 2, pp. 147-157.

104. Montgomery, D.C., 1997a. *Introduction to Statistical Quality Control,* 3-rd ed., John Wiley & Sons, New York, USA.

105. Montgomery, D.C., 1997b. *Design and Analysis of Experiments,* 4-th edition, John Wiley & Sons, New York, USA.

106. Montgomery, D.C., and Runger, G.C, 1999. *Applied Statistics and Probability for Engineers,* 2nd ed., John Wiley & Sons, New York, USA.

107. Moore, M., 2005. "Toward a Confirmatory Model of Retail Strategy Types: An Empirical Test of Miles and Snow", *Journal of Business Research,* Vol. 58, No. 5, pp. 696-704.

108. Murray, H.G., Rushton, J.P., and Paunonen, S.V., 1990. "Teacher Personality Traits and Student Instructional Ratings in Six Types of University Courses", *Journal of Educational Psychology*, Vol. 82, No. 2, pp. 250-261.

109. Neely, A., Adams, C., and Crowe, P., 2001. "The performance prism in practice", *Measuring Business Excellence,* Vol.5 No.2, pp.6-12.

110. Neely, A., and Adams, C., 2002. "Perspectives on Performance: The Performance Prism", *The Evolution of Business Performance Measurement Systems research project,* Cranfield School of Management, Great Britain.

111. Nguyen, K.D., and McInnis, C., 2002. "The possibility of Using Student Evaluations in Vietnamese Higher Education", *Quality in Higher Education,* Vol. 8, No. 2, pp. 151-158.

112. Ogata, K., and Goodkey, R. 1998. "Redefining Government performance", *Performance Measurement Publication, Alberta Finance, Government of Alberta.* Available on-line at http://www.finance.gov.ab.ca/publications/measuring/cambridge_paper.html

113. O'Neil Jr., H.F., and Bensimon, E.M., 1999. "Designing and Implementing an Academic Scorecard", *Change,* Vol.31 No.6, pp.32-41.

114. OPB, 2001. *Achieving The Oregon Shines Vision: The 2001 Benchmark Performance Report,* Report to the Legislative Assembly, Oregon Progress Board, Oregon, United States. Available on-line at: http://www.econ.state.or.us/opb/2001report/reporthome.htm

115. Owens, A.M., 2002, November 20. "Educators argue for 'real-life indicators'", *National Post,* pp. A1, A5.

116. Pang, N.S., 1996. "School values and teachers' feelings: a LISREL model", *Journal of Educational Administration,* Vol. 34, No. 2, pp.64-83.

117. Paswan, A.K., and Young, J.A., 2002. "Student Evaluation of Instructor: A Nomological Investigation Using Structural Equation Modeling", *Journal of Marketing Education,* Vol. 24, No. 3, pp. 193-202.

118. Pavlov, O.V., and Saeed, K., 2004. "A resource-based analysis of peer-to-peer technology", *Systems Dynamics Review,* Vol. 20, No. 3, pp. 237-262.

119. Perelman, L.J., 1980. "Time in System Dynamics", appears in Legasto, A.A., Forrester, J.W., and Lyneis, J.M., editors, 1980. *System Dynamics.* Studies in the Management Sciences, Vol.14, North-Holland Publishing Company, New York, NY, USA.

120. Pierre, C.B., and Mathios, D., 1995. "Statistical Process Control and Cooperative Learning Structures: A Data Assessment and Improvement System Used in a San Jose State University Summer Bridge Programme", *European Journal of Engineering Education*, Vol. 20, No. 3, pp. 377-384.

121. Pulat, B.M., 1997. *Fundamentals of Industrial Ergonomics,* 2nd ed., Waveland Press, Inc., Prospect Heights, IL, USA.

122. Quesenberry, C.P., 1991a. "SPC *Q* Charts for Start-Up Processes and Short or Long Runs", *Journal of Quality Technology,* Vol. 23 No.3, pp.213-224.

123. Quesenberry, C.P., 1991b. "SPC *Q* Charts for a Binomial Parameter *p*: Short or Long Runs", *Journal of Quality Technology,* Vol. 23 No.3, pp.239-246.

124. Quesenberry, C.P., 1995. "On Properties of *Q* Charts for Variables", Journal of Quality Technology, Vol. 27, No. 3, pp. 184-202.

125. Rauf, D., 2004. "Great Schools at Great Prices", *Careers & Colleges,* Vol. 25, No. 2, pp. 40-44.

126. Reklaitis, G.V., Ravindran, A., and Ragsdell, K.M., 1983. *Engineering Optimization, Methods and Applications,* Wiley-Interscience, John Wiley and Sons, New York, NY, USA.

127. Richmond, B., Peterson, S., and Soderquist, C., 2000. *STELLA: An Introduction to Systems Thinking,* High Performance Systems, Inc., Hanover, NH, USA.

128. Salhieh, L., and Singh, N., 2003. "A system dynamics framework for benchmarking policy analysis for a university system", *Benchmarking,* Vol. 10, No. 5, pp. 490-498.

129. Sanders, D., and Coleman, J., 2000. "Considerations Associated With Restrictions on Randomization In Industrial Experimentation", *Quality Engineering*, Vol. 12, No. 1, pp. 57-64.

130. Sanders, M.S., and McCormick, E.J., 1993. *Human Factors in Engineering and Design,* 7th ed., McGraw-Hill, Inc., New York, USA.

131. Schultz, E.M., Betebenner, D., and Ahn, M., 2004. "Hierarchical Logistic Regression in Course Placement", *Journal of Educational Measurement,* Vol. 41, No. 3, pp. 271-286.

132. SDS, 2005. "What is System Dynamics", System Dynamics Society, State University of New York at Albany, Albany, NY, USA. Available on-line at: http://www.albany.edu/cpr/sds/index.html

133. SEMNET, The Structural Equation Modeling Discussion Network, http://www2.gsu.edu/~mkteer/semnet.html.

134. Simeonov, P.L., Hsiao, H., Dotson, B.W., and Ammons, D.E., 2003. "Control and perception of balance at elevated and sloped surfaces", *Human Factors*, Vol. 45, No. 1, pp. 136-148.

135. Smith, P., 1995. "On the unintended consequences of publishing performance data in the public sector", *International Journal of Public Administration,* Vol.18 No.2&3, pp.277-310.

136. Statistics Canada, 2005. *Table 385-0001 – Consolidated federal, provincial, territorial and local government revenue and expenditure, for fiscal year ending March 31, annual (Dollars).* Available on-line at: http://www.statcan.ca/english/Pgdb/govt01b.htm.

137. Starr, P.J., 1980. "Modeling Issues in System Dynamics", appears in Legasto, A.A., Forrester, J.W., and Lyneis, J.M., editors, 1980. *System Dynamics.* Studies in the Management Sciences, Vol.14, North-Holland Publishing Company, New York, NY, USA.

138. Stiefel, L., Schwartz, A.L., and Iatarola, P., 2001. *Determinants of school performance in New York elementary schools: results and implications for resource use.* New York University, Condition Report prepared for the Educational Finance Research Consortium.

139. Sterman, J.D., 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World,* The McGraw-Hill Companies, Boston, MA, USA.

140. Tam, M., 2001. "Measuring Quality and Performance in Higher Education", Quality in Higher Education, Vol. 7, No. 1, pp.47-54.

141. Taylor, J., 2001. "Efficiency by Performance Indicators? Evidence from Australian Higher Education", *Tertiary Education and Management,* Vol. 7, No. 1, pp. 41-55.

142. Teddlie, C., and Reynolds, D., 2000. *The International Handbook of School Effectiveness Research,* Falmer Press, UK.

143. US Budget, 2005. Budget of the United States Government, Fiscal Year 2005. The Budget Documents, Department of Education. Office of Management and Budget, The Executive Office of the President of the United States. Available on-line at http://www.whitehouse.gov/omb/budget/fy2005/education.html.

144. USNews 2005. *U.S. News & World Report "America's Best Colleges 2005"*, available on-line at http://www.usnews.com/usnews/edu/college/cohome.htm.

145. Watty, K., 2003. "When will Academics Learn about Quality?", *Quality in Higher Education,* Vol. 9, No. 3, pp. 213-221.

146. Wells, P., 2004. "What Price Education?" *MacLean's,* Vol. 117, No. 36/37, pp.58-59.

147. Wholey, J.S., 1999. "Quality Control: Assessing the Accuracy and Usefulness of Performance Measurement Systems", appears in H.P. Hatry, *Performance Measurement: Getting Results,* The Urban Institute Press, Washington, D.C. (1999).

148. Wiers-Jenssen, J., Stensaker, B., and Grogaard, J.B., 2002. "Student Satisfaction: towards an empirical deconstruction of the concept", *Quality in Higher Education,* Vol. 8, No. 2, pp. 183-195.

149. Willmore, L., 2004. "Basic Education As a Human Right", *Economic Affairs,* Vol. 24, No. 4, pp. 17-21.

150. Windham, D.M., 1998. "Effectiveness Indicators in the Economic Analysis of Educational Activities", *International Journal of Educational Research*, Vol. 12, No.6, pp. 575 – 665.

151. Woodall, W.H., and Montgomery, D.C., 1999. "Research Issues and Ideas in Statistical Process Control", *Journal of Quality Technology,* Vol. 31, No. 4, pp. 376-386.

152. Worthington, A.C., 2002. "The Impact of Student Perception and Characteristics on Teaching Evaluations: a case study in finance education", *Assessment & Evaluation in Higher Education,* Vol. 27, No. 1, pp. 49 – 63.

153. Wright, S., 1921. "Correlation and Causation", *Journal of Agricultural Research,* Vol. 20, No. 7, pp. 557-585.

Appendices

Appendix I: Performance Measurement in the Public Sector

I.1. Theoretical foundations of performance measurement in the public sector

Hillison et al. (1995) provided a brief description of three theories offering a conceptual foundation for the development and use of performance measures in the public sector:

- The property-rights theory: managers of the public organizations have no individual claim on organizational performance; they are neither rewarded for good performance nor punished for poor performance. As a consequence, instead of increasing the efficiency of their organization, they pursue personal goals such as gaining prestige and individual satisfaction, and try to maximize their organization's budget instead of to improve their use of it. Performance measures motivate managers to improve the efficiency of use of their budgets.
- The public-choice theory: managers of the public organizations pursue their own interests in the form of higher pay and promotions, and this pursuit leads to risk-aversion and sub-optimal financial decisions. While the issue of self-interest is present in the private sector as well, the consequences of not meeting operational goals in the public sector are not as severe as in the private sector. Performance measures are intended to increase management's accountability.
- The principal-agent theory: the separation of the ownership from the control of an organization creates two problems: first, the goals of the agent are not consistent with the goals of the principal, and verification of the agent's actions is costly for the principal; second, the agent is more risk-averse than the principal, and the agent tends to satisfy only the minimum acceptable performance criteria. A performance measurement system creates an incentive system that encourages the agent to reveal privately held information and act

in the principal's interests. A system that prevents an agent's negative actions is considered superior to a system that only detects and penalizes underperformance.

I.2. Application of performance measurement

The task of public organizations is to provide services to the public by achieving the best possible use of the available resources. To manage their organizations effectively, executives need timely and accurate information, and their decisions should be based on facts. A performance measurement system will provide such information for the decision-making process. The following list outlines the possible uses of the performance measurement information:

- Providing managers with the information required for planning (forecasting needs and monitoring performance against targets), budgeting (quantification of budget estimates, reallocation of resources), control and decision-making (Bititci et al. 1997, Glover 1992, GAO 1980);
- Providing linkages between the performance measurement and employees' reward systems (identifying expected and achieved performance levels and linking awards to performance) (Hillison et al. 1995, Ghobadian and Ashworth 1994, GAO 1980);
- Providing a tool by which managers' performance can be assessed (makes them explain poor performance and gives them the ability to document a good performance); (Glover 1992, GAO 1980);
- Managing programs: monitoring performance, taking corrective actions, evaluating consequences of those actions (Hatry 1999, Glover 1992);
- Improving efficiency and effectiveness of managers (Ghobadian and Ashworth 1994);
- Controlling third-party providers of a service (Hillison et al. 1995);

- Finding the trade-offs between the different dimensions of performance, such as improving the program's efficiency (reducing the program's cost) while maintaining its effectiveness (Ghobadian and Ashworth 1994, Glover 1992);
- Increasing the productivity of the knowledge workers – a category to which employees of educational organizations belong (Ghobadian and Ashworth 1994);
- Measuring the intangible benefits of the programs in addition to measuring the economic benefits through the use of the financial indicators (Andersen and Fagerhaug 2002).
- Increasing organizational accountability to the public, elected officials, and higher levels of authority (Berman and Wang 2000, Foltin 1999, Hatry 1999, Ogata and Goodkey 1998, Hillison et al. 1995, Ghobadian and Ashworth 1994).

I.3. Selection of performance measures

I.3.1. Criteria for selecting measures

The following requirements can be applied to the performance measures to be selected for use (Andersen and Fagerhaug 2002, Berman and Wang 2000, Bourne et al. 2000, Foltin 1999, Hatry 1999, Alberta Finance 1996, Hillison et al. 1995, Glover 1992):

- Comprehensiveness: measures should cover all aspects of organizational performance, including possible negative consequences, but, at the same time, should not overlap with each other;
- Understandability: measures should make sense to the user;

- Validity: measures should capture in an accurate and timely manner what they are intended to measure;
- Usability: measures should be usable in the decision-making process;
- Relevance: measures should capture information that really matters;
- Reliability: captured information should be free from error and bias; other observers should be able to reach the same conclusions by using the same data set;
- Consistency: measures should be available over a period of time to make comparison possible;
- Comparability: comparing an organization's results to those of similar organizations should be possible;
- Alignment: measures should relate to an organization's mission, goals, and objectives;
- Measure ownership: an organization should have at least some influence over the measures;
- Precision: measures should be specific enough; vague or overly general measures make reporting and estimation difficult;
- Direction: measures should encourage behavior supporting the organizational strategy;
- Resistance to manipulability: indicators that are easily manipulated by the organization's staff should be avoided.

I.3.2. Quantitative versus qualitative measures

A performance measurement system should be balanced and should include different types of measures. The qualitative measures should be included as well, since not everything important is measurable, although the qualitative measures provide weaker evidence than the quantitative ones (Hatry 1999). Non-quantifiable outcomes have to be considered as well; otherwise, the easily measurable data (such as the financial data and cost indicators) will drive out the non-measurable or hard-to-measure data (the quality indicators) (Ghobadian and Ashworth 1994).

I.3.3. Number of measures

The number of performance measures used should be reasonable because too few measures decrease the accountability and possibilities for data analysis, while too many measures complicate the tracking and reporting of results. The number of the measures reported should decrease, and the reported data should become more general as the reporting level increases. The program managers should have the most detailed information, while the high-level officials and the public need a consistent summary without too many technical details. The reporting should be geared toward satisfying the data needs of the program managers, since they are responsible for the program's performance (Glover 1992).

A lack of consensus among the stakeholders about the set of performance measures may force organizations to report a large amount of performance data in the hope that the reviewers will be able to identify the important measures. This practice increases the agencies' burden during data collection, and the reviewers' burden during data analysis (Grizzle and Pettijohn 2002).

The amount of performance data that needs to be collected can be determined by using the economic cost-benefit approach. The data should be collected until the marginal benefits equal the marginal costs, where the benefits include improved organizational efficiency, and the costs include the direct costs of collecting and analyzing the data, and the indirect costs stemming from the disclosure of the performance data (Smith 1995).

A performance measurement system does not necessarily need to have a high degree of accuracy; a system should merely be able to display the negative and positive trends. Practitioners should consider the trade-off between high precision and the costs of collecting and analyzing data (Andersen and Fagerhaug 2002, Hatry 1999).

I.4. Developing a performance measurement system

Bourne et al. (2000) suggested that developing a performance measurement system has three major stages: the design, implementation, and the use of the performance measures. Each stage can be broken down even further (Andersen and Fagerhaug 2002, Foltin 1999, Hatry 1999, Glover 1992):

- The design of the performance measures
  1. Define the mission of the program (the basic objective of the program). For a mission to remain relatively unchanged, it should be stated in qualitative terms. If a mission statement is not readily available, a detailed program description may help to identify it.
  2. Define the program's stakeholders: the customers, partners, suppliers, etc. Identify both those who can benefit and those who can be hurt by the program. Identify the intermediate and the long-term customers; analyze the stakeholders' needs;
  3. Identify the program's inputs and outputs;
  4. Identify the program's outcomes that the organization seeks to achieve. In identifying outcomes, an organization should use input from the customers and partners; mapping the organization's processes will help in identifying the outcomes;
  5. Select performance indicators, which can be defined as the numerical measures of the progress toward achieving an outcome. The process maps created in the previous step can help in selecting the indicators. When

selecting measures, one should consider the consumers' demand to the product or service;

6. Establish the output and outcome targets: consider what quality attributes of a service or a product are most important to the program's customers; establish the long-term targets in addition to the short-term ones;

- The implementation of the performance measures

7. Identify data sources and collection procedures. The data can be obtained from the organization's own records, customer surveys, trained observers' ratings, and by using special equipment. Establish the data-collection frequency. Any change to the data-collection procedures/frequency should be tested on a pilot scale;
8. Select the data-separation categories. Data separation is necessary to reveal trends that can be lost in aggregation, such as differences among the high- and low-performing groups, and inequities among the customer groups;
9. Compare the performance data to the benchmarks, which could be the performance during the previous periods (i.e., the internal benchmarking that provides a relative comparison); the performance of similar organizations delivering comparable services [illegible] a comparable type of customer; a recognized standard; performance of organizations in other jurisdictions or in the private sector. The performance indicators for the different customer groups and different service-delivery methods can also be compared;
10. Analyze and report the performance data. The analysis is necessary to establish the extent to which an organization has influence on the outcomes, especially on those occurring in the distant future. Particular attention should be paid to examining and reporting the data for the indicator values that are either too high or too low compared to the targets. Such data should be reported with explanatory information.

- The use of the performance measures
    11. Bourne et al. (2000) identified two initial uses of the performance measurement data: measuring the success of the implementation of the strategy on which the measures are based and challenging and testing its validity.

I.5. Obstacles to the use of performance measurement in the public sector

Taking into account the amount of interest in performance measurement in the non-profit sector, and the relatively long history of existence of performance measurement systems, one might expect to see a movement toward their greater acceptance. However, performance measurement is still either not being used, or is being used inconsistently, or is encountering obstacles in many public organizations (de Lancer and Holzer 2001, Foltin 1999). Some of the reasons for performance measurement's lack of wide-spread use are described in details below.

I.5.1. Resistance from the management/staff

One of the major obstacles to the implementation of performance measurement systems is the resistance of the management and staff (de Lancer and Holzer 2001, Bourne et al. 2000, Foltin 1999, Hatry 1999, Hillison et al. 1995, GAO 1980). Lower management can sabotage the implementation effort through foot-dragging (Berman and Wang 2000). Several factors may cause such resistance: the perception that the performance data will be used for punishment rather than for analysis purposes; the demand for accountability where managers have only limited influence on the outcomes; the excessive attention from the higher administration and media to the negative performance information; the fear of loss of pay and jobs. The resistance might be difficult to discover, as those resisting usually try to conceal their actions.

In the field of post-secondary education, the students' evaluation of instruction is a reliable, relatively uncontaminated, and useful source of information about teaching quality. Nonetheless, the use of students' ratings of instruction is still a controversial and widely criticized practice. The criticism arises when the students' ratings and satisfaction surveys are used for faculty pay, tenure, and promotion considerations (Nguyen and McInnis 2002, Marsh 1987).

The individuals involved in performance measurement must be attracted to a performance measurement project, possibly through the use of personal incentives. Managers will become involved if they believe that performance measurement is useful to their programs (Grizzle and Pettijohn 2002).

Other solutions to the problem of resistance include allowing the program managers to have some of the authority for selecting the measures, involving the internal stakeholders, including the explanatory information with the performance report, and using punitive action only when it is clearly warranted.

I.5.2. Lack of management commitment/interest

The top management's commitment is necessary for any major undertaking, and senior management's lack of interest can seriously impede the implementation of performance measurement (Berman and Wang 2000, Glover 1992). Even when external regulations mandate the use of performance measures, they might be developed, but not actually used if an organization's managers are not committed to the process (de Lancer and Holzer 2001).

Upper management's support is also necessary to provide a sufficient level of authority to those implementing the performance measurement system (Grizzle and Pettijohn 2002).

As the development of a performance measurement system can take from anywhere between one to two years (Andersen and Fagerhaug 2002, Bourne et al. 2000), the management's interest may wane at the later stages of the system's development, and the managers' attention and commitment can be diverted to other projects. Overcoming management/staff resistance and solving IT problems may shorten the system-development time and, therefore, prevent managers' interest from waning.

I.5.3. Lack of system institutionalization

If the performance measurement system is not established as a common practice within an organization, the system either will not be used to its fullest extent or will be eventually abandoned (GAO 1980). Even when the performance measures are selected, an organization still might not use them. The implementation of a performance measurement system requires change, change creates uncertainty, and uncertainty leads to resistance (de Lancer and Holzer 2001).

Performance measurement system institutionalization requires the support of the lower managers and elected officials, the commitment of the top managers, and adequate funding (Berman and Wang 2000).

I.5.4. Absence of a process leader

Lack of a person or office to maintain and coordinate the implementation efforts can be an obstacle to adopting a performance measurement system (Glover 1992, GAO 1980). Even if a performance measurement leader exists within an organization, but the system is not adopted as an everyday management practice, the leader's departure may lead to project's discontinuation (Kaplan 2001).

I.5.5. Lack of technical capacity

Berman and Wang (2000) identified several technical capacity elements necessary for the successful implementation and productive use of a performance measurement system:

- The ability to connect inputs and outcomes
- The ability to collect data in a timely manner
- The ability to analyze data
- Sufficient information technology capabilities.

Even when the technical capacity exists, organizations need to educate the users who lack sufficient technical training about the basics of data collection and analysis. Otherwise, the performance measurement system will be too complicated for the non-technical users (Bourne et al. 2000).

I.6. Organizational performance measurement frameworks

Once an agreement on a few vital indicators is achieved, they are arranged in an organizational performance measurement framework. Their grouping is intended to reflect the prioritization of the organizational goals (i.e., the long-term versus the short-term goals, or the high-level versus the low-level goals), the separation of functions (i.e., the financial from the operational), and the different needs of the organization's multiple stakeholders.

The Balanced Scorecard developed by Kaplan and Norton in the late 1980s is probably the best-known organizational performance measurement framework. The Balanced Scorecard was initially designed for the use in private-sector companies that had realized that financial measures alone were no longer capturing the full picture of organizational performance. In addition to the financial perspective, the framework introduced customer, internal process, and learning and growth perspectives.

Initially developed for for-profit organizations, the Balanced Scorecard also proved to be valuable to non-profit and government organizations. Kaplan and Norton (2001) and Kaplan (2001) described several examples of applying the Balanced Scorecard in the non-profit and government organizations. Such organizations provide services that often are intangible, and financial measures may fail to capture these services' effects. The major modification to the Balanced Scorecard was the prioritization of the "customer's" perspective. The rationale behind such prioritization was a public organization's primary responsibility to be accountable to society. Additionally, the identity of the customers of government organizations such as regulatory and law-enforcement agencies should be analyzed carefully. The "financial" perspective of public organizations can be modified from that of profit-orientation to service-, cost-, and budget-utilization.

The recently introduced "Performance Prism" (Neely and Adams 2002, Neely et al. 2001) addresses the question "Who is the customer?" by separating the customer's identity into its "Stakeholder Satisfaction" and the "Stakeholder Contribution" perspectives. The stakeholders include not only the customers, but also other groups that might be involved in the service delivery, i.e., regulators, the local community, etc. The "Stakeholder Contribution" perspective considers the value the stakeholders deliver to the organization, and, in the case of public organizations, the funds provided by the contributors. The contributors' satisfaction is addressed by considering the "Stakeholder Satisfaction".

Several local governments developed their own performance measurement frameworks. One of the earliest and most prominent ones is the "Oregon Benchmarks" program in the State of Oregon in the United States (OPB 2001). Oregon's set of performance indicators has 90 indicators, referred to as the "benchmarks," for monitoring the state's economic, social, and environmental health.

The province of Alberta, Canada, has developed its own "Measuring Up" accountability framework (Alberta Finance 2001). In this program, the targets are set in the form of three-year business plans, and the results are reported annually. The government's financial results such as the revenue and spending as well as the assets and liabilities, are reported in the "Consolidated Financial Statements" and the "Measuring Up" report, which provide information on the core performance measures.

Appendix II: Modeling Theory

II.1. Mental models – do they work?

When we cannot (or should not) manipulate a system directly, is creating and analyzing a mental model possible?

Richmond et al. (2000) reasoned that even infants are capable of building mental models through the trial-and-error approach: infants touch a hot stove, they get burnt. As long as the relationships remain simple, mental models work well in predicting the consequences of actions. However, when the outcomes occur through a chain of events that are remote in time, the human mind is not able to interpret adequately the connections between the set of assumptions and the system's output, particularly if the system contains feedback, non-linearity, time delay, correlation and interdependencies (Jackson 2000, Richmond et al. 2000, Sterman 2000, Forrester 1968).

The following list describes some other deficiencies of mental modeling:

- The assumptions of mental models are not clearly defined, and the human mind constantly changes the assumptions and consequences of a mental model, often without even being aware of the changes (Forrester 1968, Sterman 2000);
- Mental models are hard to describe to others (Richmond et al.2000, Boland and Fowler 2000);
- The wrong model conclusions are often reached by drawing on past experience (Forrester 1968);

- Complex systems are often broken down into elements to facilitate mental analysis, and the conclusions drawn based on elements fail to account for the interaction among them (Forrester 1968, Sterman 2000);
- The information available to decision-makers is often limited, time-delayed, insufficient, ambiguous, and confounded (Sterman 2000);
- Users cannot infer about dynamic systems whose data have not yet been collected or whose behavior has not yet been observed (Sterman 2000).

II.2. System's complexity and modeling

A system's complexity depends on the number of the system's components, the relationships between the elements, the attributes of the system's elements (the number of functions that an element can perform), and the degree of the system's organization (the existence of well-established rules governing the system's behavior). As a rule, simple systems have a small number of elements and interrelations among them and are deterministic in nature, while complex systems have many interacting elements (Jackson 2000).

A system may have many elements with many interactions among them, and will still be "simple" if a small number of attributes and a high degree of organization allow the known mechanisms of behavior to be used to analyze the system's elements (Jackson 2000). For example, a school bus might be perceived as a complex system, but in fact it is not. Each element has a limited number of functions – a steering wheel is used only to steer a bus. In addition, a school bus's high degree of organization means that if the driver pushes the break pedal, the breaking system will not think whether to break or not.

A complex system may have a small number of elements, but a high number of attributes and a low level of organization would both contribute to the complexity of such a system (Jackson 2000). For example, a "student – teacher" system might seem simple, but the large number of ways in which a student may reply to a particular input and the difficulty of defining the exact behavior for a given input make this system very complex and also difficult to model.

II.3. Modeling approaches

When exploring the nature of the relationships between a system's variables is the research goal, two general modeling approaches can be used (Montgomery and Runger 1999). The "mechanistic" modeling approach employs our understanding of the physical laws defining the relationships among variables. An example would be Ohm's law of the flow of electrical current in a conductor:

$$I = \frac{U}{R}.$$

This model states that a voltage source *U* and circuit resistance *R* will determine the circuit current *I*.

The mechanistic models are useful and are widely used in the applied sciences. For example, by using a mechanistic model, one can build an electrical circuit if the characteristics of the circuit's elements and the required power output are known, or one can calculate the maximum amount of stress a welded beam can withstand given the load distribution and the beam's geometry. However, the applicability of the "mechanistic" models is limited as they can be used for solving only the systems with low complexity.

When the exact theoretical mechanism is not known or is too complex to be expressed mathematically, but the parameters explaining or influencing the phenomenon of interest are known, an "empirical" model can be used (Montgomery and Runger, 1999):

$$Y = f(X_1, X_2, \ldots, X_n)\,. \qquad \text{(II.1)}$$

In this model, the form of function f is unknown and is typically estimated by using statistical methods.

II.4. Data collection

To explore the nature of the relationship among variables in an empirical model, a researcher needs a hypothesis about the relationships among the model's variables (in the form of a set of equations) and a data set to estimate model parameters.

The data can be collected in two ways: through observational study, and through a designed experiment (Montgomery and Runger 1999). Observational study is a passive method of data collection, during which a researcher does not interfere or change the system while the data are being collected and analyzed. The data are collected either as they become available or through the analysis of historical records. In a designed experiment, a researcher makes purposeful changes to the system, collects the data, and decides what elements or causes are responsible for the observed changes in the system's performance.

II.5. Model estimation

A model is estimated by using a statistical technique, typically a regression analysis, to solve a set of equations. Since an empirical model is only an approximation of the real system, the model, typically, will not fit the data exactly. Adding an error term allows for accounting for the unexplained sources of variability in the model's parameters (Montgomery and Runger 1999, Hox and Bechger 1998):

$$\mathit{Data} = \mathit{Model} + \mathit{Error}. \tag{II.2}$$

Equation (II.2) implies that the variability in the data set is explained partially by the model itself and partially by some other parameters not known to the modeler.

Regression analysis, a statistical technique for modeling and investigating the relationship between two or more variables, is often used to estimate empirical models. An empirical regression model can be presented in the following general form:

$$Y = \mathit{constant} + aX_1 + bX_2 + \ldots + \mathit{error}. \tag{II.3}$$

The true nature of the relationship between Y and $X_1, X_2, \ldots, X_n$ remains unknown, but a regression model provides a sufficient approximation. The model above is called a "linear regression model" because Y is a linear function of the variables $X_1, X_2, \ldots, X_n$ with parameters *constant, a, b*, and … . The dependent variables X_i do not have to be linear, but can be of a more complex nature, such as X^3, $X_1 * X_2$. However, as long as Y is a linear function of the model's parameters, the regression model itself is a linear model (Montgomery and Runger, 1999).

Appendix III: Structural Equation Modeling Basics

III.1. Origins of SEM

The structural equation modeling in the conceptual form of path analysis was introduced by Sewall Wright in his groundbreaking work called "Correlation and Causation" (Wright 1921). Wright noted that in the biological sciences, a complex web of interacting and often unobservable causes produces the correlation between the characteristics of interest. While the correlation between two variables could be easily computed, the correlation was the result of only all "paths of influence" connecting the variables. Wright wanted to trace the contribution of each specific cause along a specific path toward the variation in a particular characteristic. Wright advocated combining *a priori* knowledge about the cause-and-effect relationship among certain factors with the knowledge about the correlation among the factors.

Wright expressed the relationships among the variables in the form of a path diagram with the arrows connecting the variables representing the "paths of influence." Wright used circles to indicate the unmeasured (latent) variables, rectangular or square boxes to indicate the measured (observed) variables, and double-headed arrows to indicate the covariance (or the correlation) between the variables (Hox and Bechger 1998). The same notations are used today in presenting SEM models in the path diagram form (Jöreskog and Sörbom 1996b).

Wright assigned "path coefficients" to each path in the diagram to determine each path's relative importance and developed formulas for computing these coefficients. Today, in SEM terms, the path coefficients are known as the "structural coefficients" (Jöreskog and Sörbom 1996a, Duncan 1975).

III.2. Development of the SEM into a discipline

Structural Equation Modeling emerged in the 1970s when researchers moved beyond multiple regression analysis (Duncan 1975). Until the early 1970s, the estimation of even a modest structural model was difficult and expensive because of the great number of matrix equations involved. The Structural Equation Modeling method became prominent with the advances in computing technology and introduction of the LISREL (LInear Structural RELations) software developed by Karl Jöreskog and Dag Sörbom at the Educational Testing Services in the United States (Hayduk 1987). While the original models representing the structural relationships among variables were linear, today's SEM software is capable of analyzing non-linear relationships as well (Jöreskog and Sörbom 1996b).

Today, the SEM has its own scientific journal, *Structural Equation Modeling*. Research involving SEM models appears in many academic publications, and structural equation modeling is used in such diverse fields as retail management (Moore 2005), sport management (Cunningham et al. 2005), and even religion studies (Hayduk et al. 1997).

III.3. SEM model

A SEM model is divided into two parts: a structural model specifying the causal relationships among the latent variables, and a factor (or "measurement") model specifying how the unobserved (or "latent") variables manifest themselves through their observed indicators (Hayduk 1987, Hox and Bechger 1998, Jöreskog and Sörbom 1996a, Jöreskog and Sörbom 1996b).

SEM is a versatile modeling approach. The confirmatory and exploratory factor analysis can be conducted by using the general SEM approach, with only covariances, but without causal foundations, being specified. Ordinary multiple regression analysis can be carried out as well (Jöreskog and Sörbom 1996a).

A SEM model, like any causal model, can be presented in two equivalent forms: as a picture (or a "path model"), or as a set of equations (Duncan 1975, Hayduk 1987, Jöreskog and Sörbom 1996a).

III.4. SEM model in pictorial form

A SEM model in a pictorial form (see Figure III.1) is a convenient way of presenting a model when it model has a moderate number of variables.

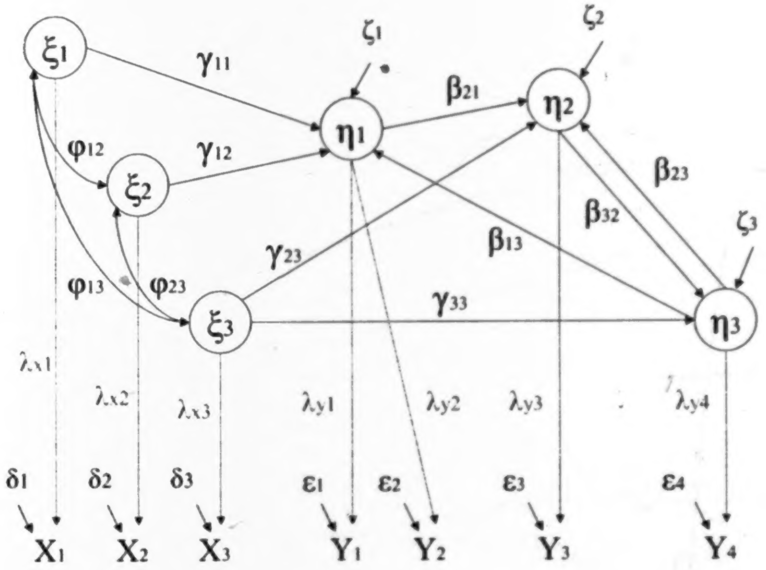

ξ1
ξ2
ξ3
η1
η2
η3
γ11
γ12
γ23
γ33
β21
β23
β32
β13
φ12
φ13
φ23
ζ1
ζ2
ζ3
λx1
λx2
λx3
λy1
λy2
λy3
λy4
δ1
δ2
δ3
ε1
ε2
ε3
ε4
X1
X2
X3
Y1
Y2
Y3
Y4

Figure III.1: SEM model in a pictorial form

In Figure III.1, LISREL's Greek notation is used to specify the model's parameters (Hayduk 1987):

- ξ_i (xi): an *exogenous* latent variable. In the model, such a variable acts only as a cause and never receives an effect (i.e., never acts as a dependent variable). We can think of these variables as the "background" variables (i.e., gender, age), or as the variables whose variation cannot be explained by the model (explaining such variables' variation is outside the model's scope);
- η_i (eta): an *endogenous* latent variable. This variable can receive effects from other endogenous and exogenous variables and can provide effects to other endogenous variables. One of the goals of research is to explain the variation in the endogenous variables;
- ϕ_i (phi): a covariance between two exogenous variables;
- ζ_i (zeta): an error variable representing the unexplained portion of variance in an endogenous variable, comparable to the error term e in ordinary regression;
- β_{ij} (beta): a structural coefficient connecting two endogenous variables. The change in the cause (one endogenous variable) by one unit is expected to directly produce β units of change in the effect (another endogenous variable). These coefficients are estimated by the model;
- γ_{ij} (gamma): a structural coefficient connecting the exogenous and endogenous variables. The change in the cause (an exogenous variable) by one unit is expected to directly produce γ units of change in the effect (an endogenous variable). These coefficients are estimated by the model;
- x_i: an indicator of a corresponding exogenous variable;
- y_i: an indicator of a corresponding endogenous variable;
- δ_i (delta), ϵ_i (epsilon): the measurement errors of the corresponding indicators. A measurement error higher than zero indicates that the latent variable will not predict the observed variable perfectly (e.g., the indicator is not a perfect measure of the underlying concept).

- λ_{Xi}, λ_{Yi} (lambda): the factor loadings. These are the structural effect coefficients linking specific indicators to the specific concepts. Specific lambda values are used to set the scale on which the values of the underlying concepts are measured.

The SEM model in Figure III.1 illustrates the additional powers this approach gives to a researcher: the separation of the unobserved theoretical concepts from their observed indicators, the possibility of the direct specification of the degree of measurement error in the indicators, and the possibility of modeling the reciprocal and inter-dependent relationships among the theoretical concepts.

III.5. SEM model in equation form

When a model has a large number of variables, a path diagram becomes cluttered. In such a case, an algebraic representation of a model is more convenient (Jöreskog and Sörbom 1996a).

A SEM model can be expressed in three basic equations. Employing LISREL's Greek notation again, the first of equations is (Hayduk 1987)

$$\underset{(m\times 1)}{\eta} = \underset{(m\times m)}{\mathrm{B}}\ \underset{(m\times 1)}{\eta} + \underset{(m\times n)}{\Gamma}\ \underset{(n\times 1)}{\xi} + \underset{(m\times 1)}{\zeta} . \qquad \text{(III.1)}$$

Equation (III.1) describes the direct effects among the concepts (latent variables). In this equation, η is a vector of m endogenous variables, B and Γ are the matrices of the structural effect coefficients, ξ is a vector of n exogenous variables, and ζ is a vector of m error terms.

The other two equations are the equations of the measurement structure:

$$\underset{(px1)}{y} = \underset{(pxm)}{\Lambda_y} \underset{(mx1)}{\eta} + \underset{(px1)}{\varepsilon}, \tag{III.2}$$

$$\underset{(qx1)}{x} = \underset{(qxn)}{\Lambda x} \underset{(nx1)}{\xi} + \underset{(qx1)}{\delta}. \tag{III.3}$$

In equations (III.2) and (III.3), the parameters *x* and *y* are the vectors of *p* observed endogenous indicators and *q* observed exogenous indicators, correspondingly. The Λ_Y and Λ_X parameters are the matrices of the structural effect coefficients, and ϵ and δ are the vectors of the measurement errors.

The four matrices of the structural coefficients (B, Γ, Λ_Y and Λ_X) and the four covariance matrices (Φ – the matrix of the covariances among the exogenous concepts, Ψ – the matrix of the covariances among the errors ζ, Θ_ϵ – the matrix of the covariances among the errors ϵ and Θ_δ – the matrix of the covariances among the errors δ) create a framework that can represent many models (Hayduk 1987).

III.6. Model estimation

A model's parameters are estimated by fitting a model to the relevant data. The data source required for a SEM model is a matrix of covariances among the observed indicators, or a so-called "matrix *S*." The use of a covariance matrix instead of the original observations provides a number of advantages to a researcher. Firstly, the amount of data required for the analysis is greatly reduced. Instead of analyzing hundreds or thousands of individual observations, a researcher can analyze a covariance matrix whose size is determined by the number of variables (numbering in tens) (Jöreskog and Sörbom 1996a). Secondly, when the privacy of the respondents is at stake, the original responses can be removed as soon as the summary covariance matrix is computed.

On the other hand, the SEM model implies that the variances and covariances among the indicators x and y are based on the known (or estimated) values of the elements of the matrices B, Γ, Λ_Y, Λ_X, Φ, Ψ, Θ_ϵ, and Θ_δ. The SEM model's coefficients are estimated by using the maximum likelihood estimation, which is carried out by comparing the matrix S and the model-implied matrix of the covariances among the observed indicators (the so-called Σ matrix) (the details of the estimation can be found in Chapter 5 of Hayduk 1987).

III.7. Assessment of the adequacy of a model's fit

The adequacy of a model can also be assessed by comparing the matrices S and Σ by utilizing the χ^2 (chi-square) function with the number of degrees of freedom equal to

$$d.f. = \tfrac{1}{2}\,[(p + q)(p + q + 1)] - t. \qquad \text{(III.4)}$$

In equation (III.4), p is the number of the endogenous variables, q is the number of the exogenous variables, and t is the number of the estimated coefficients. As the number of estimated coefficients increases (the increase can be achieved by leaving coefficients free to vary), the number of the model's degrees of freedom decreases.

Since the difference between the matrices S and Σ may arise due to the sampling fluctuations in the matrix S (if the matrix Σ is a true population matrix), the fit between the two matrices is estimated based on the P value returned by the chi-square test (Hayduk 1987).

Appendix IV: Systems Thinking

IV.1. Systems Thinking approach

When human life emerged on the Earth, the systems were not studied. The humans just accepted the systems as they were and adjusted themselves to the systems around them. The development of industrial societies created the need to understand the systems – industrial, social, biological and political. However, much of systems analysis concentrated mere system description, and analysis was mostly verbal and qualitative. Such an approach was not sufficient to understand the real nature of systems (Forrester 1968).

Revealing the structure – the interrelationships between the elements – of any field of knowledge is necessary for learning and mastering the process of understanding. At the same time, understanding is much easier when an area of knowledge is structured and systematized. The physical systems were the first to be given a structure of principles because these systems are much simpler than the social and biological systems. Studying physical systems has become more efficient as their structures have been extensively studied and explained, and the relationships among the physical elements of such systems have been established. Mathematical formulae describe a system's principles more clearly than the sets of data values in tables (think of $E = I*R$ in electric circuit versus a table of E, I, and R values), and allow for the generalization of physical variables and processes (Richmond et al. 2000, Forrester 1968).

Old learning concepts, based on knowledge accumulation, are no longer sufficient to resolve the problems that the modern world presents. Progress has brought changes in technology that have increased our power to influence the environment, but at the same time, the problems have become more serious, the risks more significant, and the impacts of our decisions and actions more profound. As the complexity of the problems increases, the consequences of actions become harder to predict. Since the power of an individual is growing as well, foreseeing not only the direct and intended results of actions, but also the indirect and unintended, becomes paramount.

The goal of learning processes today is not knowledge and data accumulation – our rapid technological progress quickly renders knowledge obsolete, and data in themselves are useless if they cannot help in making inferences about the real world – but the ability to apply knowledge and to analyze the data to reveal and understand the causal links among the elements making up a system.

"Systems Thinking" is a general term used to denote the theories, methodologies, models, tools, and techniques based on systems ideas, concepts and approaches. Systems Thinking appeared as a result of the inability of the natural sciences to solve the problems of the modern complex world. The methods of the natural sciences employed "reductionism" – breaking complex problems down into small parts, and analyzing them separately. Systems Thinking, instead, encourages the analysis of complex problems as a whole with particular attention to the interconnections and relationships among the elements and the dynamic behavior these interconnections and relationships produce. The modern physical and social sciences are moving toward a holistic view of natural phenomena, as compared to the traditional reductionist view. Acceptance of the systems view is essential for solving today's problems, whether physical or social (Jackson 2000).

IV.2. Development of Systems Thinking into a cross-discipline

Systems Thinking emerged in the 1940s and 1950s as a "transdiscipline" or "cross-discipline" based on the development of systems ideas in various social and natural sciences. Systems Thinking is not considered a "discipline" by itself, since it does not focus on a particular area of the real world for study (Jackson 2000).

The roots of Systems Thinking can be traced to the following disciplines (Jackson 2000):

- Philosophy. Aristotle in ancient Greece described the functions of an organism as a whole: the body parts cannot function by themselves (i.e., to be able to see, an eye has to be a part of the organism). Aristotle applied the same reasoning when describing the relationship of an individual and a state – the individual can fulfill her/his purpose only as a part of the state. Among the philosophers of the Western school, Spinoza considered the universe to be a single entity and that breaking it down was irrational. Several other branches of philosophy influenced the development of systems ideas and systems thinking.
- Biology. A living organism is a complex system that has a well-defined boundary separating it from the environment and that can sustain itself via transactions across the boundary. Some characteristics of an organism cannot be broken down into parts. Biologist Ludwig von Bertalanffy is described as one of the founding fathers of the transdiscipline of systems thinking. Von Bertalanffy considered a living organism as a "whole" made up of interacting and interdependent parts, and described organisms as "open systems" with such characteristics as "regulation " and "feedback." Open systems take material, energy, and information from the environment to maintain themselves in a "steady state."
- Sociology. The powerful notion of society as a system dominated the development of much of traditional sociology. Sociologists Spencer (1820-1903) and Durkheim (1858-1917) were central figures in developing a notion of society as an organism, where society is a system made up of interdependent parts working to maintain the whole. Analysis of the parts and

institutions of a social system should concentrate on their contribution to the whole.

- Management and Organization Theory. Among different models of management, systems theory eventually gained dominance over the other theories. The main deficiency of other theories was their concentration on analyzing the parts of an organization, in contrast to the systems view of an organization as a whole system composed of interrelated components. The Systems Thinking approach also advocated considering organizations as open systems constantly interacting with their environments, in contrast to other theories' view of organizations as closed systems. Comparing an organization to an organism allowed researchers to regard an organization's main goals as survival and continuous existence.
- Cybernetics. Cybernetics appeared as a separate transdiscipline during the Second World War when scientists from different fields were brought together to work on problems of communication and control in various processes. Norbert Wiener introduced the term *cybernetics* in 1947. For the control of a mechanical or a biological system, the negative feedback is of crucial importance. All behavior directed toward the achievement of a goal depends on negative feedback. Communication is another important notion, since control involves the communication of information. Forrester's System Dynamics originated from management cybernetics. Forrester's idea was that the behavior of complex systems could be analyzed by modeling their dynamic feedback processes.

Systems appropriate for cybernetic study (versus statistical or operational research) possess the characteristics of extreme complexity, self-regulation, and probabilism. Cybernetics has tools to deal with each of these characteristics:

  - *Extreme complexity* – is dealt with by using the "black box" technique. A "black box" is a complex system that cannot be easily analyzed to determine the nature of the processes governing its behavior. Instead of analyzing a "black box" directly, the experimenter should manipulate the inputs and observe the outputs in order to find the patterns (regularities) of its behavior. If nothing is known about the system, random inputs can be used as well. As knowledge is gained about the system's behavior, more structured experiments can be conducted.

- *Self-regulation* – self-regulation is a desirable characteristic of a complex system that can be achieved via negative feedback mechanisms. Managers of complex organizations can best achieve stability and pursue desired goals by introducing appropriate feedback processes.
- *Probabilism* – complex systems are probabilistic. They exhibit a great variety of states of behavior. To control a system, we need as much variety as a system has. If a system's variety exceeds our capabilities to respond, the ways to control the system is either to reduce its variety or to increase our own capabilities.

Although Systems Thinking originated from different disciplines, systems research and practice can be both multi- and mono-disciplinary. Research can be multi-disciplinary in the sense of involving people from different disciplines with different views of a problem that might come from a particular field of science.

IV.3. The Structure of Systems Thinking

The structure of contemporary systems thought can be divided into three broad categories (Jackson 2000):

- Discipline-based Systems Thinking: research of systems ideas in separate disciplines such as biology, sociology, philosophy, and cybernetics;

- Study of systems themselves: the goal research in this category is to develop a General Theory of Systems;
- Applied Systems Thinking: the application of systems thinking for problem solving with the orientation on the "problem-owner."

Theoretically, the applied Systems Thinking approaches can be classified according to the relevant types of underlying social theory (see Figure IV.1).

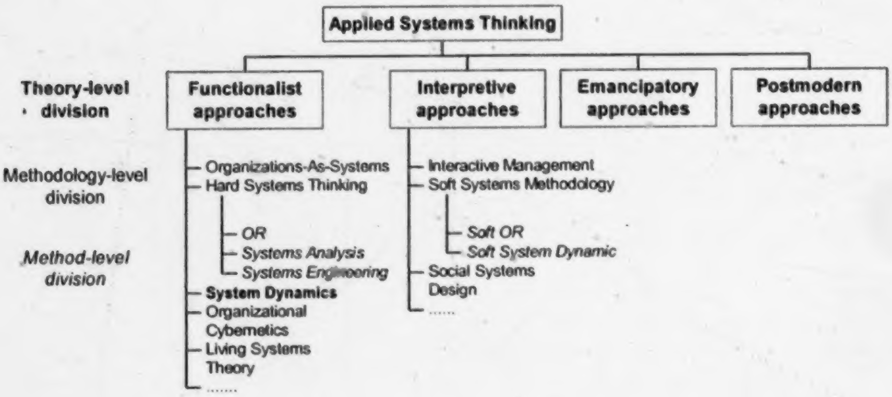

Applied Systems Thinking
Theory-level division
Functionalist approaches
Interpretive approaches
Emancipatory approaches
Postmodern approaches
Methodology-level division
Method-level division
Organizations-As-Systems
Hard Systems Thinking
OR
Systems Analysis
Systems Engineering
System Dynamics
Organizational Cybernetics
Living Systems Theory
....
Interactive Management
Soft Systems Methodology
Soft OR
Soft System Dynamic
Social Systems Design
....

Figure IV.1: Systems Thinking theoretical approaches

These four generic social paradigms are based on the assumptions social scientists make about the nature of social science and society, and provide "frameworks of ideas" for systems approaches (Jackson 2000):

- Functionalist systems approach: systems are assumed to exist as real objects independent from us as observers. Scientific methods are used to discover the relationships among the elements of a system. The knowledge gained during such research is used to improve a system's efficiency, adaptability, and survivability. Humans can be treated as elements of a system, and their presence in the system does not require a change in the principles of system analysis. Social systems can be regulated to achieve their optimal level of performance. The main goal of the functionalist systems approach is the prediction and control of mechanical and social systems.

- Interpretive systems approach: humans are assumed to have priority over the system's technology, structure, or organization. Different individuals and groups within a social system frequently have different views of what course of action a system should pursue. People possess free will, and their points of view and intentions have to be understood if we want to change and improve the system of which they are a part.
- Emancipatory systems approach: in this approach, a society is seen as oppressing some groups being discriminated against on the basis of gender, class, age, or other characteristics. Empowering or emancipating those currently suffering should radically reform the present social order. The critique of a current social order has to be combined with action to overcome "false consciousness" of the oppressed so they can see reality and change the system. Emansipatory ideas may seem alien to the systems field, but during the 1980s and 1990s, a number of researchers sought to expand the social role of systems thinking by using the critical tradition in philosophy and social theory. The goal of the approach is to create a better social system via radical change.
- Postmodern systems approach: "the postmodern approach seeks, through methods such as deconstruction and genealogy, to reclaim conflict and to ensure that marginalized voices are recognized and heard" (Jackson 2000, p. 117).

IV.4. Inability of Hard Systems methods to resolve dynamic problems

Hard systems thinking is a collection of applied methods including Operations Research, Mathematical Optimization, Systems Analysis, and Systems Engineering. The assumption in applying methods of hard systems thinking is that a clearly identifiable goal is known beforehand. Four assumptions are made about a problem (Jackson 2000):

1. The desired system state is known;
2. The present system state is known;
3. Alternative methods can be used to transform the present state into the desired one;
4. The analyst's goal is to find the best way of transforming the present state into the desired state by evaluating alternatives in terms of performance measurement.

Hard Systems Thinking methods are appropriate for systems that have a clearly defined goal (final state) and when the performance metric against which to judge the system's effectiveness is available. Hard systems approaches yield satisfactory results for engineering-type problems, where organizations can be treated as machines governed by identifiable rules and laws. However, the produced solutions will be unreliable and distorted when such tools are used on systems possessing the following qualities (Legasto and Maciariello, 1980, Sterman 2000):

- Outcomes occur through a chain of events,
- Outcomes are remote in time,
- Actions may produce unforeseen undesirable consequences,
- Relationships among a system's elements involve feedback, non-linearity, time delay, correlation and/or interdependencies.

Hard Systems Thinking methods fail so solve complex dynamic problems due to several limitations (Jackson 2000):

- Limited applicability: only problems with clearly defined objectives can be dealt with successfully. In situations where the objectives are not agreed upon, and every stakeholder has its own interests and values, the hard systems thinking approach has to consider one set of objectives as superior to others, even though doing so is not always correct.
- Deterministic approach to human behavior: human behavior is modeled mathematically like the other mechanical parts of a system. The fact that humans support change only if it has a favorable meaning for them is ignored.
- Complex systems cannot be easily modeled mathematically: factors that cannot be easily quantified are sometimes ignored.
- Hidden power conflict: in complex situations, favoring one set of objectives, usually that of the most powerful stakeholder, becomes necessary.

Appendix V: Application for Study Approval

University of Alberta

Research Ethics Committee

Application for Study Approval

Student: Kostyantyn Grygoryev	**Faculty:** Stanislav Karapetrovic

Study Title:
SQC Modeling of Teaching and Learning in a Classroom

Study Description:

As a part of his research on a Ph.D. thesis entitled "An Integrated Model for Performance Measurement, Modeling and Management in Education", Mr. Kostyantyn Grygoryev, a Ph.D. Candidate in the Department of Mechanical Engineering will perform a study on using Statistical Quality Control (SQC) to measure and monitor teaching and learning performance in a classroom setting. The study will be conducted during the Winter 2003 semester in sections B1 and B2 of the ENGG 401 Fundamentals of Engineering Management course in the Departmental of Mechanical Engineering at the University of Alberta. The study consists of the following three phases:

- Mr. Grygoryev designs a total of twenty-eight (28) to thirty-two (32) multiple-choice questions regarding the material covered in the ENGG401 course, in collaboration with the two instructors of the course, Mr. Jay Cameron and Dr. Stanislav Karapetrovic. A total of eight (8) to nine (9) questionnaires, each containing four (4) questions are compiled. An example of a questionnaire that will be used is provided in Appendix A of this application.
- Mr. Grygoryev administers the questionnaires in eight (8) to nine (9) classes in both sections, according to the agreement with the course instructors. Informed consent is asked from all students at the time the first questionnaire is administered. The students are also informed that the participation in the study is completely voluntary and anonymous, and that the purpose of the study is to improve course delivery and learning outcomes. The students answer the questions before and after the class, without identifying their names, student numbers or any other personally-identifiable information. Both filled and empty questionnaires are placed in boxes by the classroom entrance doors.
- Mr. Grygoryev collects the statistics of the proportion of correct answers after the lecture, as well as the proportion of incorrect answers before and correct answers after the lecture. These statistics are plotted on appropriate p-charts, and out-of-control conditions are monitored throughout the course. Follow-up actions, such as the repetition of a part of a lecture or provision of additional examples may be suggested.

Study Benefits:

This study is aimed at improving the teaching process and learning outcomes in a classroom setting. The instructor benefits from the continuous monitoring of his/her performance and student learning outcomes and the identification of possible problems early in the course. The students benefit from the focus on several most important points in each lecture, and hence improved learning outcomes. The researchers benefit from the ability to validate the proposed SQC model in a real-life setting.

Study Risks:

No specific risks to people are expected from this research study.

Ethical Considerations:

Informed consent:

The consent will be asked from all students before the first questionnaire is administered, by reading the statement enclosed in Appendix B of this application. Participation in the study is completely voluntary and anonymous. The filling of answers (i.e. A, B, C or D) in the appropriate boxes of the questionnaire (see Appendix A) will constitute student consent.

Anonymity:

The study is completely anonymous. The students will be asked not to record any personally-identifiable data (for example names or student numbers) on the questionnaires. If a student writes his/her name or student number by mistake, this will be erased immediately after the summation of data has begun.

Other aspects:

No deception and/or concealment will be deployed in this research. No potentially hazardous equipment and/or material will be used in this research.

ANNEX A: Sample Questionnaire

ENGG 401 BEFORE AND AFTER LECTURE QUESTIONS (Jan. 30, 2003)

1. In the example with three companies A, B and C presented in class, which of the three companies is best off in the short term?
   a. Company A
   b. Company B
   c. Company C
   d. I don't know

ANSWER: BEFORE ☐ AFTER ☐

2. Which of the following statements about break-even points is TRUE?
   a. Book break even is the point at which operating income plus depreciation becomes >0
   b. Cash break even is the point at which operating income becomes >0
   c. A small business owner is more focused on cash break even than on book break even
   d. I don't know

ANSWER: BEFORE ☐ AFTER ☐

3. Which of the following statements about the level payment loan is TRUE?
   a. The amount paid for interest decreases with time
   b. The amount paid for interest increases with time
   c. The amount paid for interest does not change with time
   d. I don't know

ANSWER: BEFORE ☐ AFTER ☐

4. Which of the following entries from the income statement is included on the statement of retained earnings?
   e. SG&A
   f. Net Income
   g. COGS
   h. I don't know

ANSWER: BEFORE ☐ AFTER ☐

ANNEX B: Research Study Information

Study Title:
SQC Modeling of Teaching and Learning in a Classroom

Research Investigators:

Kostyantyn (Kosta) Grygoryev	Stanislav (Stan) Karapetrovic
Office: TEMP LABS 1-40	Department of Mechanical Engineering
Department of Mechanical Engineering	University of Alberta
University of Alberta	T6G 2G8 Edmonton, Alberta
kg6@ualberta.ca	S.Karapetrovic@ualberta.ca
Phone: (780) 492-9734	(780) 492-9734

Research Description:

Good morning. My name is Kostyantyn Grygoryev. I am conducting a study on how to use statistical quality control in measuring, monitoring and improving lecture delivery and learning outcomes. This research is a part of my doctoral research in the Department of Mechanical Engineering at the University of Alberta. The purpose of the study is to improve the teaching process and overall teaching and learning performance. In about eight to nine lectures during this course, I will be distributing questionnaires regarding the course material covered in each of those lectures. Each questionnaire will contain four questions with multiple-choice answers. You may answer these questions before and after each lecture in which the questionnaire is administered. You are under no obligation to participate in this study. The participation is completely voluntary, and your choice whether to participate or not will bear no consequences or effects on your mark in this course. This study is completely anonymous. If you choose to participate, please do not write any personally-identifiable information, such as your name or student number, on the questionnaire. Please leave your questionnaires in the designated box by the classroom entrance. If you decide to participate, your written answers will constitute your written consent to participate in this study. If you decide not to participate, you may leave your empty questionnaire in the designated box. If you have any questions regarding this study, please do not hesitate to contact me, or the study coordinator Dr. Stanislav Karapetrovic. Any questions regarding the ethical considerations in conjunction with this study should be directed to Dr. John Whittaker, department of Mechanical Engineering, University of Alberta.

Appendix VI: Before and After questions appearing in MBKP sets 1, 2, and 3

1. You have just been hired as a manager of a computer store that buys computers directly from the manufacturers and then sells them to retail customers. Which of the following existing accounts are you most likely to close in order to increase efficiency and save accounting costs in the store:

   a. Payables
   b. Receivables
   c. WIP
   d. COGS
   e. I don't know

2. My company has just paid a supplier $2,000 for the material that was received a month ago. Which of the following statements is true?

   a. Payables go up by $2,000 and cash goes up by $2,000
   b. Finished Goods Inventory goes up by $2,000 and receivables go up by $2,000
   c. WIP goes up by $2,000 and cash goes up by $2,000
   d. Payables go down by $2,000 and cash goes down by $2,000
   e. I don't know

3. Which of the following periods is the LEAST likely to be covered by an income statement:

   a. Daily
   b. Monthly
   c. Quarterly
   d. Yearly
   e. I don't know

4. The last line ("bottomline") on an income statements is commonly:

   a. COGS
   b. Net Income
   c. Revenue
   d. Contibution Margin
   e. I don't know

5. The statement that the contribution margin is the same as net income is:

a. True
b. False
c. I don't know

6. Which of the following statements is TRUE?

a. When starting a new business, annual pro-forma statements are required for two years maximum
b. Year to year consistency in statements is less important than using a standard form
c. Manufacturers often have a significant COGS that must be tracked to keep costs down or prices keep up with the rising cost
d. I don't know

7. If the contribution margin is falling, it is often a sign of:

a. Decreasing COGS and constant prices
b. Sales department not doing enough to extract price from the customers
c. Returns, recalls and other quality problems
d. Both answers (b) and (c) are correct
e. I don't know

8. Which of the following statements is FALSE?

a. Contribution margins can vary widely from product to product even in the same company
b. Margin can be a crucial measure in identifying how well a business is doing year to year
c. COGS can be almost zero for some businesses
d. Margin should never be expressed as a percentage of gross revenue
e. I don't know

9. If my contribution margin in 1999 was $500,000, in 2000 it was $400,000 and in 2001 it was $200,000, it may be a sign of:

a. Business going well, because my contribution to the costs is decreasing
b. Business not going well, because my COGS is probably increasing
c. Business not going well, because my COGS is decreasing and my prices are increasing
d. I don't know

10. In theory, which of the following statements is TRUE?

a. COGS is a fixed expense, SG&A is a variable expense
b. COGS is a fixed expense, SG&A is a fixed expense
c. COGS is a variable expense, SG&A is a fixed expense
d. I don't have a clue

11. If your company has a very small cost (e.g. of supply material) that is variable, and you do not measure this cost per product or per unit of labor time, you are most likely to put this cost into:

a. COGS
b. SG&A
c. Allowance for warranty and returns
d. I don't know

12. Which of the following statements is TRUE?

a. By definition, SG&A expenses is calculated per product or per product line
b. The salary of the President of a company is normally put as a COGS
c. Property taxes that a company pays are normally put into SG&A
d. I don't have a clue

13. In the example with three companies A, B and C presented in class, which of the three companies is best off in the short term?

a. Company A
b. Company B
c. Company C
d. I don't know

14. Which of the following statements about break-even points is TRUE?

a. Book break even is the point at which operating income plus depreciation becomes >0
b. Cash break even is the point at which operating income becomes >0
c. A small business owner is more focused on cash break even than on book break even
d. I don't know

15. Which of the following statements about the level payment loan is TRUE?

a. The amount paid for interest decreases with time
b. The amount paid for interest increases with time
c. The amount paid for interest does not change with time
d. I don't know

16. Which of the following entries from the income statement is included on the statement of retained earnings?

a. SG&A
b. Net Income
c. COGS
d. I don't know

17. The statement of cash flow includes the following three types of activities:

a. Operating, Financing and Costing
b. Operating, Investing and Financing
c. Financing, Cashflowing and Maintaining
d. I don't have a clue

18. Which of the following statements about the Statement of Cash Flow is TRUE?

a. Amortization is included in the operating activities
b. Long-term debt is included in the investing activities
c. Dividends paid are included in the maintaining activities
d. I don't know

19. If my net income is (+)$100K, and the total of my depreciation and change in non-cash working capital is (+)$50K, which of the following statements is TRUE?

a. Cash flow from operations is $150K
b. Cash flow from operations is $50K
c. Cash flow from all activities is $150K
d. I have no idea

20. If my balance sheets show that my receivables have decreased by $20K, what will be the impact of this on the Cash Flow Statement?

a. There will be no impact
b. Cash flow will increase by $20K
c. Cash flow will decrease by $20K
d. I don't know

For the following three questions, refer to the financial statements of Dofasco for the year 2001 (Provided before the class).

21. The "Quick Ratio" or "Acid Test" is:

- a. 2.85
- **b. 1.06**
- c. 0.48
- d. I don't know

22. The "Days Sales Outstanding Ratio" is:

- **a. 38**
- b. 46
- c. 103
- d. I don't know

23. The "Debt Ratio" is:

- a. 18.3%
- b. 13.5%
- **c. 31.9%**
- d. I don't know

24. Which one of the following ratios is the most important for owners / investors in a company?

- a. Times Interest Earned
- **b. Return on Equity**
- c. Total Asset Turnover
- d. I don't know

25. Which of the following concepts is used to reduce work-in-progress inventory:

a. Just Any Time (JAT)
b. Just in Time (JIT)
c. Total Quality Management (TQM)
d. I don't know

26. The three corners of the "balancing triangle" in financial management are:

a. Market share, margin and cost
b. Market share, income and cost
c. Margin, income and revenue
d. I don't know

27. The "Tragedy of the Commons" example shows that:

a. Commodities have cyclical prices
b. Inflation is the only reason why commodity prices tend to go up
c. Commodity producers generally act for common, rather than individual interest
d. I don't know

28. The fundamental truth about the financial management of a company that I learned in this class can be summed up as:

a. "Operating income is breath"
b. "Clean books make a good sale of a company"
c. "Cash is breath"
d. I don't know

Appendix VII: Data Collected From MBKPs

Set 1
Instructor A
Winter 2002 semester

Question number	Sample size, n(i)	Number of IBCA	Number of IA
1	69	24	4
2	69	6	1
3	69	1	0
4	69	3	21
5	65	16	8
6	65	26	22
7	65	18	23
8	65	17	35
9	64	12	4
10	64	15	3
11	64	25	10
12	64	14	11
13	57	28	3
14	57	6	30
15	57	22	16
16	57	14	4
17	47	5	5
18	47	11	24
19	47	6	19
20	47	7	30
21	44	21	0
22	44	22	7
23	44	28	8
24	44	14	6
25	57	19	1
26	57	30	5
27	57	29	4
28	57	6	1

Set 2
Instructor A
Fall 2002 semester

Question number	Sample size, n(i)	Number of IBCA	Number of IA
1	81	2	9
2	81	1	1
3	81	3	0
4	81	10	2
5	56	9	3
6	56	9	16
7	56	12	19
8	56	13	23
9	63	7	14
10	63	30	22
11	63	47	10
12	63	14	9
13	48	37	1
14	48	2	25
15	48	24	11
16	48	3	5
17	42	15	3
18	42	12	17
19	42	5	21
20	42	7	21
21	38	20	1
22	38	21	1
23	38	18	13
24	38	16	9
25	43	13	5
26	43	14	15
27	43	20	10
28	43	5	3

Set 3
Instructor B
Fall 2002 semester

Question number	Sample size, n(i)	Number of IBCA	Number of IA
1	102	17	35
2	102	7	9
3	102	5	14
4	102	23	18
5	101	17	10
6	101	12	24
7	101	14	33
8	101	8	42
9	104	23	21
10	104	25	44
11	104	43	50
12	104	10	17
13	95	11	36
14	95	3	50
15	95	3	75
16	95	23	18
17	92	52	6
18	92	17	64
19	92	15	45
20	92	16	54
21	80	20	13
22	80	20	25
23	80	13	43
24	80	16	18
25	85	6	12
26	85	10	36
27	85	38	5
28	85	7	3

Set 4
Instructor A
Winter 2003 semester

Question number	Sample size, n(i)	Number of IBCA	Number of IA
1	88	18	19
2	88	16	8
3	88	29	13
4	88	4	17
5	74	7	18
6	74	27	12
7	74	22	14
8	74	13	41
9	74	9	5
10	74	5	6
11	74	8	28
12	74	8	5
13	60	20	5
14	60	14	7
15	60	5	6
16	60	8	17
17	62	31	4
18	62	19	20
19	62	11	14
20	62	4	14
21	55	7	10
22	55	7	[illegible]
23	55	7	[illegible]
24	55	9	13
25*	46	16	19
26*	46	8	3
27*	46	12	26
28*	46	27	12

*** - B&A questions in set 4 different from corresponding questions in sets 1, 2, and 3**

Appendix VIII: Overview of Statistical Quality Control

VIII.1. Brief history of statistical quality control

Quality has always been an important characteristic of any product or service. However, formal methods of scientifically controlling quality evolved over time. The principles of scientific management were developed in the early 1900s, and in 1924, Walter A. Shewhart introduced a statistical control chart. Statistical acceptance sampling, as an alternative to 100% inspection, was developed by the end of the 1920s at the Bell Telephone Laboratories, but World War II drove the expansion and acceptance of statistical quality control (SQC) methods in the industry.

At the end of World War II, Japan used the statistical principles of process control and improvement to rebuild the devastated Japanese industries. Japan's success was such that the U.S. automobile industry was almost destroyed in the 1980s by Japanese competitors offering cheaper, more reliable cars than the American models. Japanese companies were able to excel by systematically applying SQC methods in every aspect of a product's lifecycle – from new product development, to the evaluation of the design, to manufacturing, to field performance. The global market forced the United States to re-discover statistical control methods in the 1980s when, for many companies, quality became a matter of survival. Statistical quality control methods were introduced into American industry and academia. The US chemical companies' current strong position in the world market is credited to their early adoption of SQC tools. Today, "Total Quality Management," the "Six Sigma," and "ISO 9000" are household names. These are the management systems within which statistical quality control methods are only a part of a quality-driven business improvement philosophy (Montgomery 1997a, Woodall and Montgomery 1999).

VIII.2. Statistical Control Chart

A statistical control chart is the most sophisticated tool in the arsenal of problem-solving tools employed to control processes methodically and scientifically. The premise behind this chart is that to meet the customer's requirements, a product must be produced by a stable process, and that variability is inversely proportional to quality (Montgomery 1997a).

The following example illustrates the reasoning behind this premise. Imagine a manufacturing line packaging 500-gram cans of coffee (as per customer requirement). If the company puts more than 500 grams of coffee in a can, the company will loose money (even though customers probably will not mind); and if the company puts less than 500 grams of coffee in each can, the company will lose customers.

Statistical process control is used to reduce the variability of a process by eliminating assignable causes of variation (such as an ill-adjusted machine or operator error), leaving only the chance causes of variation (inherent, or natural, variability) in the process. A statistical control chart is a tool of statistical process control and is used to detect the occurrence of assignable causes and eliminate them before a large amount of non-conforming output is produced (Montgomery 1997a).

A statistical control chart displays a quality characteristic computed from a sample over time or a sample number. The chart has a center line representing the process' average value, and two control limits, the upper and lower limits (not the same as the "specification bounds"). Control limits are typically set at plus/minus three sample mean standard deviations. A statistical control chart, in effect, tests the hypothesis that the process mean is equal to its target value, at different points in time (Montgomery 1997a) (see Figure VIII.1).

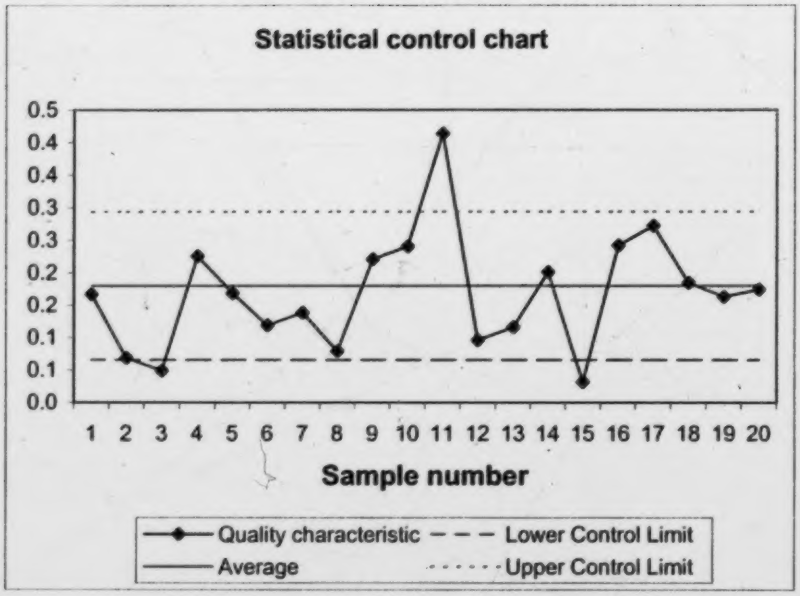


Figure VIII.1: Example of a statistical control chart

If a point (i.e., the calculated sample statistic) plots inside the control limits, the process is considered to be in the state of statistical control, and attempts to adjust the process' parameters will only make it worse; therefore, when in statistical control, the process should be left to continue to operate unchanged.

In his famous experiment, Dr. W. Edwards Deming, one of the founders of statistical quality control, dropped a marble through a funnel centered at the bull's eye of a target and plotted the distance from the marble to the center of target. The experiment was run for the first time by dropping marbles without changing the position of the funnel, and for the second time by trying to compensate for the error by moving the funnel for an equal and opposite distance of the error recorded on the previous drop. When the funnel remained untouched, the variance of the distance of a marble from the target center was about twice less than in the case when funnel was moved to adjust for the error. This results happened because the error from the previous drop provided no information on what the error on the next drop would be – an error was truly random, and the process was operating under the presence of random causes only (Montgomery and Runger 1999).

When a point plots outside the control limits, an investigation is necessary to determine whether a process shift occurred and an assignable cause of variation is present or if a random cause produced the observed out-of-control situation. On this chart, points 3, 11, and 15 indicate an out-of-control situation, which should trigger a search for assignable causes. When assignable causes are eliminated, a process is said to be in "a state of statistical control" (Montgomery 1997a). The in-statistical-control process still might produce non-conforming output, however. If no assignable causes of variation are evident in a process producing 490-grams cans of coffee, we still have a problem. The statistical quality control task was completed when the non-random causes of variation were removed. The further adjustments to the process are a matter of process design and quality management.

Appendix IX: Before and After Questions Appearing in "Variable" MBKPs

1. Briefly (in several sentences) describe the difference between quality control and quality management.

2. You are planning to purchase a new laptop computer. Briefly (in several words) describe each of the quality characteristics (quality dimensions) of a computer that might be important in making a selection.

3. In some credit card companies, departments issuing credit cards and departments collecting debt are managed and rewarded separately. The separation creates an incentive for the credit-issuing department to issue as many credit cards as possible, so their performance would look good. Credit cards are being issued to people with questionable credit history, which increases the amount of uncollectable debt. This worsens performance of the debt-collecting department. Briefly explain which quality principle is being overlooked in such company?

4. Name five differences between ISO 9000: 1994 series and ISO 9000: 2000 series

5. What is the purpose of a gap analysis in implementing a quality management system?

6. In developing the generic quality management standards for worldwide application, ISO/TC 176 drew on a considerable national experience of several countries with well-established quality management standards. Can you name those countries and national standards?

7. Briefly describe the concept of quality loop flowcharting.

8. What benefit(s) will ISO 10018 provide to consumers? To businesses?

9. Briefly describe the difference between quality inspection and quality audit.

10. Briefly define 1st - party, 2nd - party, and 3rd - party audit.

11. What is the difference in underlying models between the quality-related standards (ISO 9000 series) and social responsibility standards (ISO 14000, OHSAS 18000)?

12. Name two components required for a successful implementation of an Integrated Management System.

13. What is the relationship between Six Sigma and Process Capability?

14. What is the major difference in quality assurance principle between the ISO 9001 standard and Business Excellence Models (such as European Quality Award, Malcolm Baldrige National Quality Award) ?

15. Taking into account the amount of interest to the performance measurement and time that performance measurement systems have been around we might expect that there is a movement toward its greater and greater acceptance. But performance measurement is still not being used, used inconsistently, or encounters several obstacles in many organizations. What do you think those obstacles are?

16. One of the requirements/criteria that performance indicators should satisfy is *relevance*: measures should capture information that really matters. What other requirements/criteria can you think of?

Appendix X: Variable Knowledge Transfer Statistic, Raw Data

Question	1			2			3			4			5		
Answers	B	A	KG	B	A	KG	B	A	KG	B	A	KG	B	A	KG
1	9	10	1	5	9	4	7	7	0	5	6	1	8	9	1
2	7	8	1	3	3	0	8	8	0	2	4	2	10	10	0
3	0	3	3	2	3	1	6	6	0	1	4	3	5	9	4
4	6	7	1	2	4	2	5	5	0	4	9	5	8	10	2
5	4	7	3	2	3	1	0	5	5	3	3	0	9	10	1
6	0	5	5	2	2	0	9	9	0	4	7	3	9	9	0
7	5	9	4	1	4	3	2	2	0	3	3	0	5	8	3
8	7	8	1	3	3	0	5	8	3	4	6	2	7	8	1
9	5	7	2	5	6	1	6	7	1	3	5	2	2	9	7
10	8	10	2	5	8	3	8	8	0	5	6	1	2	8	6
11	6	8	2	4	6	2	6	6	0	6	6	0	9	9	0
12	6	6	0	5	5	0	2	2	0	0	2	2	6	7	1
13	1	5	4				1	1	0	4	6	2	4	8	4
14	1	4	3	5	5	0	5	5	0	2	4	2	3	9	6
15	1	4	3	3	4	1	6	6	0	0	2	2	1	1	0
16	5	7	2	2	5	3	5	6	1	4	8	4	7	10	3
17	7	9	2	4	5	1	7	8	1	1	10	9	4	8	4
18	4	7	3	3	7	4	6	7	1	1	2	1	6	6	0
19	6	7	1	5	6	1	8	8	0	6	9	3	4	8	4
20	5	5	0	4	7	3	2	2	0	0	1	1	5	8	3
21	7	7	0	5	7	2	9	10	1	10	10	0	6	9	3
22	5	7	2	3	6	3	2	2	0	4	7	3	0	7	7
23	6	6	0	5	5	0	6	6	0	9	10	1	6	9	3
24	5	9	4	3	3	0	3	3	0				8	10	2
25	6	9	3	3	7	4							5	9	4
26	7	7	0	2	2	0							5	6	1
27	8	8	0	4	4	0							5	6	1
28	5	6	1	3	3	0									
29	5	8	3	3	6	3									
30	5	6	1	2	5	3									
31	5	8	3	5	7	2									
32	7	7	0	3	3	0									
33	9	10	1	3	3	0									
34	2	7	5	4	9	5									
35	7	9	2	4	4	0									
36	7	9	2												
Sample size			36			34			24			23			27
Average			1.94			1.53			0.54			2.13			2.63
Variance			2.11			2.38			1.39			3.94			4.70
st dev			1.45			1.54			1.18			1.98			2.17

B – Before A – After KG – Knowledge Gain

6			7			8			9			10			11		
B	A	KG	B	A	KG	B	A	KG	B	A	KG	B	A	KG	B	A	KG
4	8	4	5	7	2	0	8	8	5	7	2	5	6	1	1	10	9
4	8	4	4	8	4	5	5	0	7	7	0	6	6	0	1	1	0
8	9	1	7	8	1	5	8	3	7	7	0	9	10	1	1	1	0
6	10	4	5	6	1	0	3	3	3	3	0	8	8	0	1	10	9
9	9	0	8	9	1	1	3	2	2	2	0	6	9	3	1	1	0
5	9	4	5	5	0	10	10	0	7	8	1	5	5	0	1	1	0
3	7	4	6	7	1	5	6	1	7	10	3	8	10	2	1	10	9
5	7	2	3	3	0	4	4	0	6	6	0	4	6	2	2	10	8
3	8	5	4	6	2	2	2	0	1	2	1	5	7	2	1	1	0
8	9	1	7	7	0	7	7	0	8	10	2	9	9	0	2	2	0
0	7	7	6	8	2	7	7	0	5	5	0	5	8	3			
0	8	8	8	9	1	8	8	0	7	7	0	7	7	0			
0	8	8	7	7	0	7	7	0	5	5	0	7	7	0			
4	7	3	4	7	3	2	2	0	6	7	1	1	3	2			
5	9	4	5	6	1				5	5	0	10	10	0			
0	1	1	8	10	2				8	10	2	6	6	0			
5	9	4	8	8	0				6	6	0	10	10	0			
0	8	8	5	6	1				5	7	2	1	1	0			
0	8	8	5	7	2				6	7	1	4	7	3			
9	9	0	7	8	1				4	5	1	5	7	2			
8	8	0	4	6	2				6	6	0	4	4	0			
8	8	0	4	7	3				7	7	0	1	2	1			
8	9	1	5	8	3				5	6	1						
5	8	3	8	9	1												
0	2	2	0	8	8												
6	9	3															
		26			25			14			23			22			10
		3.42			1.68			1.21			0.74			1.00			3.50
		7.05			2.89			5.10			0.84			1.33			20.9
		2.66			1.70			2.26			0.92			1.15			4.53

12			13			14			15			16		
B	A	KG	B	A	KG	B	A	KG	B	A	KG	B	A	KG
3	7	4	1	1	0	2	3	1	2	2	0	2	2	0
1	2	1	9	9	0	2	3	1	2	3	1	3	6	3
1	1	0	3	3	0	0	2	2	4	5	1	3	3	0
1	10	9	2	2	0	3	4	1	1	5	4	2	2	0
10	10	0	2	2	0	2	3	1	4	10	6	2	7	5
1	10	9	4	4	0	2	3	1	8	8	0	2	2	0
2	2	0	4	4	0	0	3	3	1	1	0	3	7	4
1	1	0	2	2	0	3	5	2	4	4	0	0	1	1
1	10	9				3	4	1	2	2	0			
5	10	5				2	3	1						
						4	8	4						
						1	3	2						
						2	2	0						
		10			8			13			9			8
		3.70			0.00			1.54			1.33			1.63
		16.46			0.00			1.10			4.75			4.27
		4.06			0.00			1.05			2.18			2.07

Appendix XI: Use of Designed Experiments to Discover Assignable Causes

XI.1. Strategy for testing the "poor question" or "poor lecture" hypothesis

To investigate whether a poor "Before and After" question resulted in students' poor performance on a MBKP, several variations of questions that were deemed ambiguous can be tested during the next semester that the course is scheduled. Several variations of an MBKP including different variations of a suspected question should be prepared, but the same lecture outline and style should be used.

Depending on how many students are present during a lecture covering the subject of a question, several variations of the "before and after questionnaire" should be randomly distributed. One variation should involve the original question, and another one (or two, or three – depending on how many students are attending a lecture, to maintain a reasonable sample size) should involve modified questionnaires. In the Design of Experiments terms, this approach is called a "completely randomized single-factor design" (Montgomery 1997b). The question will be the factor, and the question type (original, modified 1, modified 2,…) will be the level. If the number of students is not equal in each group (a likely case, for even when we distribute an equal number of questionnaires of each type, the response rate has always been less than 100%), we will have an *unbalanced* design.

To test whether a poor lecture was the reason behind a students' poor performanceon a MBKP, deliver a modified lecture in the next semester and compare the results with those of the previous semester.

XI.2. Data analysis

The following table (Table XI.1) can be used for recording the results of a "single-factor design" (question only, or lecture only) experiment:

Table XI.1: Results of completely randomized single-factor design (adopted from Montgomery 1997b)

Question type (levels)	Students' Answers											Totals
Original	Y_{11}	Y_{12}	…	…	…	…	…	…	…	…	Y_{1ni}	$Y_{1.}$
Modified 1	Y_{21}	Y_{22}	…	…	…	…	…	…	…	…	Y_{2ni}	$Y_{2.}$
Modified 2	Y_{31}	Y_{32}	…	…	…	…	…	…	Y_{3ni}			$Y_{3.}$
....	…	…	…	…	…	…	…	…	…	…	…	…
a	Y_{a1}	Y_{a2}	…	…	…	…	…	…	…	Y_{ani}		$Y_{a.}$
										Grand total		$Y_{..}$

In this table, Y_{11} is a correct/incorrect answer to the original question by Student 1 from the group which received the original question. If Y_{11} is coded on a "0 – incorrect, 1 – correct" scale, then $Y_{1.}$ is the total number of correct or incorrect answers computed for a student group, and $Y_{..}$ is the grand total of correct/incorrect answers in the class. Note that the number, and not the proportion, is computed in this case.

The statistical model of this design will be (Montgomery, 1997b)

$$Y_{ij} = mean + M_i + error_{ij}, \qquad \text{(XI.1)}$$

where M_i is the effect of the *i*-th question type, and the error is assumed to be randomly and independently distributed with the mean equal to zero. If the original question (lecture) is compared to a single alternative, a standard *t* test on the difference between two means can be used. If more than two means are compared, the consecutive comparison of pairs of means will lead to an increase in "type I" error (rejecting the hypothesis of the equality of means, when the hypothesis is true). The proper way to test equality of several means is by employing the analysis of variance method (Montgomery 1997b).

The ANOVA (ANalysis Of VAriance) test procedure is presented in Table XI.2.

Table XI.2: ANOVA table, completely randomized single-factor design (adopted from Montgomery 1997b)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F - test
Factor	$SS_F = \sum_{i=1}^{n_i} \frac{Y_{i.}^2}{n_i} - \frac{Y_{..}^2}{N}$	$a-1$	$MS_F = \frac{SS_F}{a-1}$	$F_{a-1,N-a} = \frac{MS_F}{MS_{error}}$
Error	$SS_{error} = SS_{total} - SS_F$	$N-a$	$MS_{error} = \frac{SS_{error}}{N-a}$	
Total	$SS_{total} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij}^2 - \frac{Y_{..}^2}{N}$	$N-1$		

In this table, when testing different questions, n_i is the number of students returning questionnaire i, a is the number of types of a question, and N is the total number of observations (total number of students returning questionnaires).

XI.3. Strategy for testing "poor question and poor lecture" hypothesis

When both poor lecture and question quality are suspected as assignable causes, a factorial experiment is necessary. The statistical design will have two factors – the lecture and the question and will need at least two levels of each factor – the old design and the new design. The factorial arrangement of a two factors-two levels experiment is presented in Table XI.3:

Table XI.3: Factorial design for testing "poor lecture and poor question" hypothesis

		Question	
		Old	New
Lecture	Old	P_{IA} P_{IBCA}	P_{IA} P_{IBCA}
	New	P_{IA} P_{IBCA}	P_{IA} P_{IBCA}

A class would be randomly divided into two groups. One group would be given an old lecture and two types of each question. Another group would be given a new lecture and two types of each question.

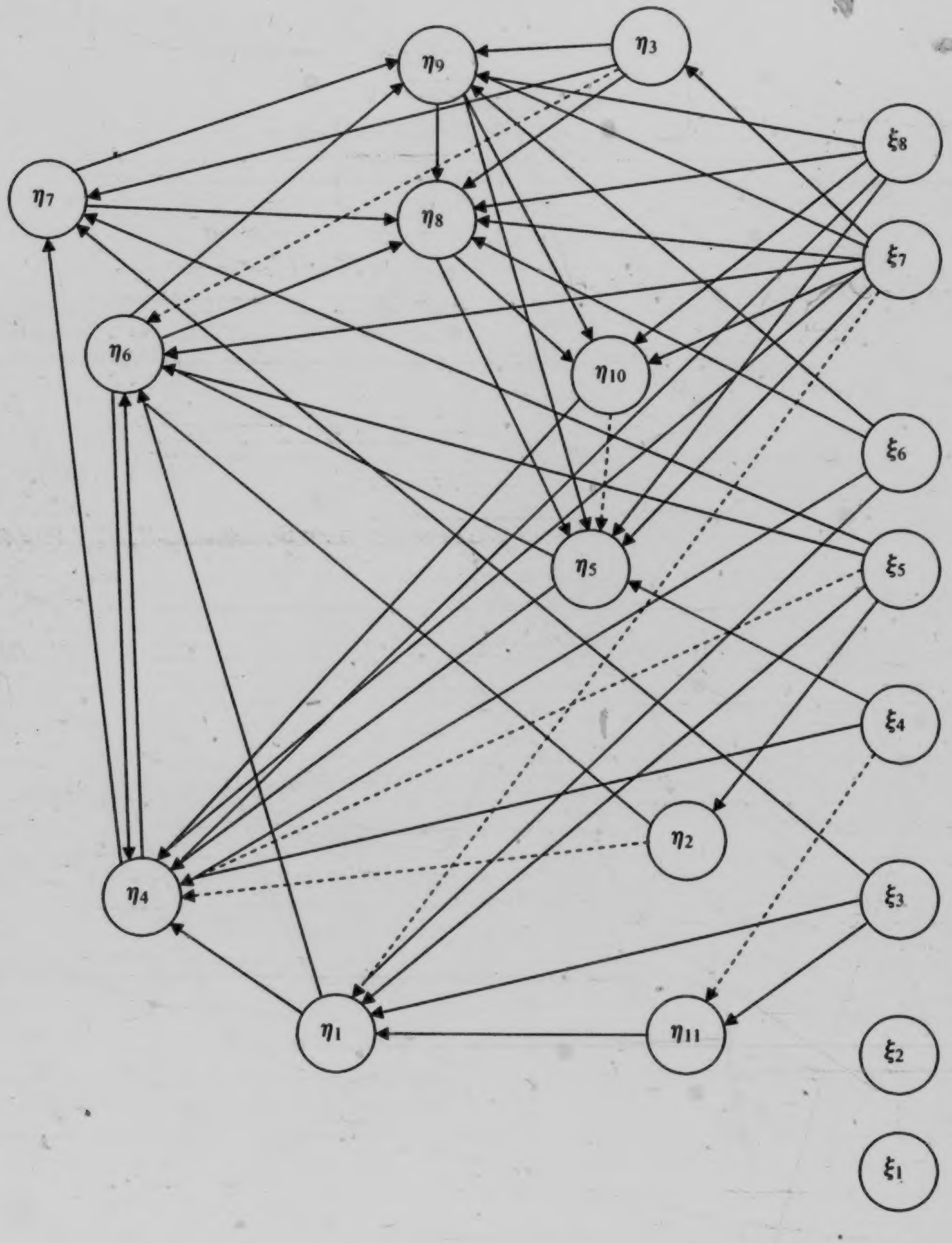
Appendix XII: Matrix of Covariances Among Observed Indicators (Matrix S), SEM Classroom Educational Model

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12	X1
EXTRACTV	1.0241												
SPEKENGL	0.1239	0.8284											
SMLRCRSE	0.0285	-0.0009	0.5579										
READTEXT	-0.1305	-0.0756	-0.0106	0.7487									
HWRKTIME	-0.2097	-0.0868	-0.0175	0.3026	0.7270								
LTRSMISS	0.0385	-0.0089	0.0103	-0.1246	-0.1500	0.5397							
MIDTSCRE	0.1396	0.1532	-0.0016	0.1169	0.0584	-0.1284	1.3971						
WORKLOAD	-0.0805	-0.0899	-0.0567	0.0564	0.1716	-0.0422	-0.0501	0.5832					
SATISFCT	0.0732	0.0586	0.0262	0.1680	0.1101	-0.1475	0.5003	-0.1215	1.3060				
INTRGJOB	0.0132	0.0555	-0.0423	0.1512	0.1477	-0.1457	0.2464	-0.0970	0.9035	1.4069			
LKESBJCT	-0.0487	0.0320	-0.0036	0.1218	0.1308	-0.1044	0.2586	-0.0745	0.6868	0.6937	1.1362		
DOWLHVFN	0.1007	0.0195	-0.0061	-0.0069	-0.0247	-0.0013	0.0052	-0.0095	-0.0035	-0.0200	-0.0308	0.1389	
OWNACOMP	-0.0183	-0.0029	0.0129	0.0048	0.0066	0.0067	-0.0155	-0.0039	0.0043	0.0106	0.0004	-0.0022	0.0563
STUDTSEX	0.0119	0.0054	-0.0108	0.0264	0.0320	-0.0234	0.0506	0.0158	-0.0329	-0.0283	-0.0340	0.0082	-0.0033
HAVEFUNU	0.0137	0.0030	0.0004	0.0057	-0.0050	-0.0100	0.0109	0.0083	0.0086	0.0100	0.0060	0.0030	0.0018
DOWELLUN	-0.0191	-0.0025	-0.0024	0.0437	0.0200	-0.0409	0.0396	0.0079	-0.0085	-0.0132	-0.0004	0.0013	-0.0007
LANGBGND	-0.1546	-0.1477	0.0390	0.1262	0.1569	-0.0654	-0.0435	0.1043	0.1121	0.1303	0.1371	-0.0161	0.0039
STUDTAGE	-0.1524	-0.0372	0.0304	0.1241	0.1429	-0.0425	0.0433	0.0537	0.1598	0.1571	0.1573	-0.0035	0.0016
GRADLEVL	-0.0566	-0.0425	0.0275	0.0771	0.0840	-0.0444	0.0479	0.0362	0.0948	0.0923	0.0780	0.0009	0.0007
INSTEXPR	-0.0155	-0.0724	0.0398	0.1965	0.1864	-0.1554	0.1680	-0.0662	0.2957	0.3300	0.1884	-0.0043	-0.0010

	X2	X3	X4	X5	X6	X7	X8
EXTRACTV							
SPEKENGL							
SMLRCRSE							
READTEXT							
HWRKTIME							
LTRSMISS							
MIDTSCRE							
WORKLOAD							
SATISFCT							
INTRGJOB							
LKESBJCT							
DOWLHVFN							
OWNACOMP							
STUDTSEX	0.1634						
HAVEFUNU	0.0027	0.0129					
DOWELLUN	0.0113	0.0045	0.0517				
LANGBGND	0.0033	0.0016	0.0074	0.4331			
STUDTAGE	0.0103	0.0003	0.0040	0.1269	0.3118		
GRADLEVL	0.0083	0.0013	0.0054	0.0902	0.0808	0.0892	
INSTEXPR	0.0345	0.0039	0.0203	0.1042	0.0730	0.1093	1.1506

Appendix XIII: SEM Models "Original" and "Final" in a Path-Diagram Form

- - - - - - - - - - - -▸ Structural effects present in the "Original" model and removed from the "Final" model



η_3
η_9
ξ_8
η_7
η_8
ξ_7
η_6
η_{10}
ξ_6
η_5
ξ_5
ξ_4
η_2
η_4
ξ_3
η_1
η_{11}
ξ_2
ξ_1

Appendix XIV: LISREL Syntax, Model "Original"

```
DA NI=20 NO=384 MA=CM

!Data: Number of Input variables: 20, Number of Observations: 384, Matrix to Analyze: Covariance
Matrix

CM sy

***Matrix of covariances among observed indicators (Matrix S, Appendix 4) entered here***

LA

!indicators' labels

EXTRACTV SPEKENGL SMLRCRSE READTEXT HWRKTIME LTRSMISS
MIDTSCRE WORKLOAD SATISFCT INTRGJOB LIKESBJCT DOWLHVFN OWNACOMP
STUDSEX HAVEFUNU DOWELLUN LANGBGND STUDTAGE GRADLEVL INSTEXPR

MO NY=12 NX=8 NE=11 NK=8 LY=FU,FI LX=ID BE=FU,FI GA=FU,FI PH=FU,FR PS=DI,FR
TE=DI,FI TD=DI,FI
!Model specs
! NY-number of Y indicators, NX-numb of X indicators, NE-num of eta variables, NK-num of ksi vars

! Lambda X is identity matrices since there is a single indicator per concept
! Lambda Y is full fixed (at 0), and values 1 will be assigned to single indicators,
! and one of multiple indicators for concept ETA 4 ("Time to course") will be freed
! Beta and Gamma are full, and fixed (at initial values 0). Later the coefficients to be
!estimated will be freed

!PSI is the matrix of covariances among the errors of exogenous vars ETA. It is diagonal to exclude
!covariances between ETAs' errors (as they are assumed to be independent). The matrix is free, so that
LISREL would estimate error variances

! Tay E and Tay D are the matrices of covariances between the errors of indicators. These matrices are
!diagonal to exclude possible covariancesbetween indicators' errors.

FR BE(1,11), BE(4,1), BE(4,2), BE(4,5),BE(4,6), BE(4,10), BE(5,8), BE(5,9), BE(5,10)
FR BE(6,1), BE(6,2), BE(6,3), BE(6,4), BE(6,5), BE(7,3), BE(7,4), BE(8,3), BE(8,6), BE(8,7),
BE(8,9)
FR BE(9,3), BE(9,6), BE(9,7), BE(10,8),BE(10,9)
! free up beta elements that will be estimated by LISREL


FR GA(1,3), GA(1,5), GA(1,6), GA(1,7),GA(2,5), GA(3,7), GA(4,4), GA(4,5), GA(4,6), GA(4,7),
GA(4,8)
FR GA(5,4), GA(5,7), GA(5,8), GA(6,5),GA(6,7), GA(7,3), GA(7,5), GA(8,6), GA(8,7), GA(8,8)
FR GA(9,6), GA(9,7), GA(9,8), GA(10,7), GA(10,8), GA(11,3), GA(11,4)

VA 1.0 LY(1,1), LY(2,2), LY(3,3), LY(54), LY(6,5), LY(7,6), LY(8,7), LY(9,8)
VA 1.0 LY(10,9), LY(11,10), LY(12,11)

FR LY(4,4)
!Multiple (second) indicator of the concept "Time to course"
```

```
VA 0.1024 TE(1,1)
! 10% error var to the indicator "extra activities"
VA 0.16568 TE(2,2)
! 20%
VA 0.02789 TE(3,3)
! 5%

FR TE(4,4)

VA 0.3635 TE(5,5)

VA 0.10793 TE(6,6)
! 20% to Lectures missed
VA 0.13971 TE(7,7)
! 10% to Test Score
VA 0.05832 TE(8,8)
! 10% to Workload
VA 0.26119 TE(9,9)
! 20% to Satisfaction with course
VA 0.28137 TE(10,10)
! 20% to Instructor did a good job
VA 0.22724 TE(11,11)
! 20% to Like subject
VA 0.01388 TE(12,12)
! 10%
VA 0.00281 TD(1,1)
! 5%
VA 0.00163 TD(2,2)
! 1%
VA 0.00064 TD(3,3)
! 5%
VA 0.00258 TD(4,4)
! 5%
VA 0.04331 TD(5,5)
! 10%
VA 0.03118 TD(6,6)
! 10%
VA 0.00089 TD(7,7)
! 1%
VA 0.23012 TD(8,8)
! 20%

OU ME=ML ALL ND=4
!output: method of estimation-max likelihood, numb of digits-4
```

Appendix XV: System Dynamics Modeling Steps

XV.1. Key variables

The two approaches used in system dynamics to identify model-relevant relationships and factors are empirical and theoretical data analysis, and use of interviews with decision-makers to identify the rules by which decisions are made. Through observation, discussions, and the examination of past decisions and their consequences, an analyst who understands the dynamic effects of feedback structures can separate the factors that led to the decisions, but nothing guarantees that two equally proficient analysts will arrive at the same problem formulation. Building a model – defining its purpose, selecting variables, deciding on the validation approach – is as much art as science, and no definite rules on modeling exist. The particular approach to model building depends on a researcher's vision. (Legasto and Maciariello 1980, Starr 1980).

XV.2. Time horizon

The time horizon should be relevant for the system under study, depending on the model's purpose – whether it is used for forecasting, scenario generation, or policy analysis (Sterman 2000, Legasto and Maciariello 1980). A model's time horizon is, essentially, a function of the modeler's interests (Perelman 1980).

The temporal boundary of a system dynamics model can be analyzed by using the Newtonian mechanic concept and thermodynamic concept (Perelman 1980). Time as Newtonian concept can be reversed, in a way so that inserting *t* in an equation with a negative sign will predict motion in the opposite direction, and a system dynamics model can be run in reverse to produce the same results backwards. The thermodynamic concept assumes that time in unidirectional and follows the Entropy Law. The real world exhibits thermodynamic time behavior: transformation in the universe is irrevocable because of the degradation of heat, friction, and other phenomena. While the theoretical discussion of system dynamics is thermodynamic, the programming of models is Newtonian. The thermodynamic concept of time should be considered superior since it provides a more realistic context.

XV.3. Model boundary

A model's purpose is a statement of the goal of the modeling effort (Richmond et al. 2000). In defining a model's purpose, a researcher must decide what should be left outside the model, and how the real world can be simplified (Richmond et al. 2000, Sterman 2000). Recall that the SEM approach advocates achieving model parsimony, or fitting data with the fewest number of estimated structural coefficients (Hayduk 1987).

A model boundary should separate all relevant elements responsible for the dynamic behavior of a system under study from the rest of the world. All the interactions considered to be important should lie within the boundary, and the only interaction allowable across the boundary is a random disturbance (Jackson 2000, Richmond et al. 2000, Bell and Senge 1980, Legasto and Maciariello 1980, Starr 1980, Forrester 1968).

In defining the model boundary, a researcher is not disregarding the many other forces influencing the system, but is trying to explain it by using the simplest possible set of relationships. In SD, a model should include the smallest possible number of elements generating observed dynamic behavior (Richmond et al. 2000, Forrester 1968).

A parsimonious (generic) model is applicable to a wider range of cases and has greater chances of being refuted or corroborated by each new data set (Bell and Senge 1980, Hayduk 1987).

Different views of the model boundary will exist for the same problem, since each analyst is guided by her/his own perception of the world, and this perception is related to the analyst's values, attitudes, and academic background. Trying to reconcile views coming from two different disciplines will mean subjecting one of the disciplines (which can be SD) to another discipline's standards. Nonetheless, the ideas of the participating analysts should not be ignored, and even when the scope of a model is being defined, no *a priori* limit should be set on the model boundary (and time horizon). The different model boundaries can be tested through simulations. (Legasto and Maciariello 1980).

XV.4. Development of dynamic hypotheses

During this step, a dynamic hypothesis for a problem is developed by identifying the system's feedback and stock-and-flow structure. A dynamic hypothesis can be first expressed verbally (Richmond et al. 2000): "Students spending more time on studying improve their test score, but, at the same time, the workload may become excessive, which will negatively affect grades." Then the verbal hypothesis is transcribed into a causal loop diagram and, eventually, into a stock-and-flow map (Sterman 2000).

Several approaches can be used to develop a causal diagram and stock-and-flow map. A modeler may look for the "key actor" (Richmond et al. 2000) or the most dominant feedback structure (Sterman 2000). Another approach is to model the main flow (for example, the main flow of knowledge) in the system (Richmond et al. 2000).

XV.5. Model formulation

At this stage, the model is translated into its equation form, and the model's parameters and initial conditions are specified. Writing down equations helps in detecting the formulations flaws and inconsistencies, which are present when a modeler has to resort to "fudge factors" to maintain dimensional consistency (Sterman 2000).

A system can be described in integral and differential forms. Integration represents the accumulation of the flows in the stocks, while differentiation represents measuring the speed of the flows (e.g., represents the rates). Forrester (1968) argued in favor of using integral rather than differential equations to describe a system. Integration is the natural process of physical and biological systems, while differentiation never occurs in the real world. Differentiation involves measuring velocity instantaneously, but no natural or artificial device can do so. Describing a system in terms of differential equations disconnects the world of mathematics from the real world and may even suggest the reverse of factual causality relationships. Expressing the flow as a derivative of the level implies that a change in the level creates a flow, but, indeed, the level changes because of the flow (Forrester 1980). Thus, rather than expressing a system in the terms

$$\text{Rate} = d(\text{level})/dt,$$

we should use equations such as:

$$\text{Level} = \int(\text{Rate})\, dt.$$

XV.6. Model testing and model validity

The goal of modeling is not to build the "right" model. Models will differ from reality since a model is necessarily a simplification of reality and relies on imperfectly measured data. Because we do not have perfect information about anything in the physical or social world, we cannot absolutely prove that a model is correct. At the same time, we have at least partial knowledge about everything. A model should be judged on its ability to improve our understanding of a system and to describe a system better than we could do mentally and verbally. The goal of modeling is to develop a clear representation and definitions that can be easily communicated, and model validity is of relative value since the quality of a mathematical model should be evaluated against the mental model we would use otherwise (Sterman 2000, Legasto and Maciariello 1980, Forrester and Senge 1980, Forrester 1968).

Instead of asking "Is the model right?" a modeler should ask, "Is the model useful?" Validation, therefore, is the process of establishing confidence in a model's quality and usefulness, and the ultimate goal of validation is to convey this confidence to the model's users. A model's validity should be assessed relative to the model's purpose, since all models are inaccurate to some degree, and all of them can be falsified by some test (Sterman 2000, Forrester and Senge 1980).

In SEM the model fit is assessed by using a chi-square test, and a number of diagnostics based on the statistical properties of a sample are provided in the model's output. In SD, the statistical testing of a system dynamics model's output is only an element in the overall model validation. Passing a statistical test is not a sufficient condition of model validity, as the probability of a Type II error can be high, and failure to pass a statistical test should not be considered as a reason for rejecting a model. A model displaying no significant inconsistencies with the whole range of real-world data can be deemed "valid" even if the results from individual tests are "weak." A model is compared to empirical evidence to corroborate or refute a model, and empirical information is not limited to numerical statistics alone (Forrester and Senge 1980, Legasto and Maciariello 1980, Starr 1980).

The purpose of a SD approach is learning and describing instead of predicting and prescribing. A "valid" model, therefore, should produce behavior *qualitatively* coherent with the real world, and "perhaps consistent with an accepted statistical treatment of aggregated numerical data" (Starr 1980).

No single test exists to validate a system dynamics model. The more tests a model passes, the higher is the degree of confidence in the model's usefulness and quality (Forrester and Senge 1980). Several tests for system dynamics models are summarized in the Tables XIV.1 and XIV.2 below (Sterman 2000, Forrester and Senge 1980):

Table XV.1: SD model tests of structure (based on Sterman 2000, and Forrester and Senge 1980)

Test	Purpose	Methodology
Structure verification	Comparing the model's structure with the structure of the real system being modeled	Review of the model's assumptions first by a modeler, then by people experienced with the real system; comparing the model's assumptions to the real system's decision-making, and the literature.
Parameter verification	Comparing the model's parameters (constants) with the parameters of the real system being modeled	Conceptual correspondence – parameters should exist in a real system; numerical correspondence – parameter values should have a realistic range of values. If parameter is likely to change its value over the simulation time and policy regions, it should be converted into a variable with an associated structure.
Extreme conditions	Evaluating the model's behavior under the extreme combinations of levels (state variables) – minus infinity, zero, plus infinity	Each rate (policy) equation should be tracked to the level (state) on which the rate depends, and the reasonableness of the resulting rate equations should be evaluated (e.g., if inventories are zero, shipments must be zero too).
Boundary adequacy (structure)	Evaluating the model's aggregation levels, and inclusion of all the relevant structure	A hypothesis relating a new model structure with a particular issue should be identified. If the importance of feedback interaction between a model and new structure in dealing with the issue cannot be demonstrated, the

		model passes the test[2]. The model's boundary should not be expanded to include structures irrelevant to the particular purpose.
Dimensional consistency	Analyzing the dimensional consistency of the rate equations	The necessity of including the "scaling" parameters having no real-life meaning indicates a problematic model structure.
Statistical tests		Variables should not be excluded from a model based solely on the statistical insignificance of the parameter estimates (t-values)

---

[2] An example is the *Urban Dynamics* model: the new model structure is the suburbs, and a particular issue is the ineffectiveness of job-training programs. If the importance of city-suburb interactions in the effectiveness of job-training cannot be demonstrated, the model passes the boundary adequacy test (Forrester and Senge 1980).

Table XV.2: SD model tests of behaviour (based on Sterman 2000, and Forrester and Senge 1980)

Test	**Purpose**	**Methodology**
Behavior reproduction tests	A group of tests examining the match between the model's behavior generated through the model's *internal* policies and the real system's historical behavior. Observed behavior should be a result of the model's structure.	*Symptom generation* – a model's policies and structure should reproduce the symptoms (problems) that led to the model's creation; *Multiple mode test* – the ability of a model to generate more than one mode of behavior observed in the real world (for example, short- and long-term fluctuations in employment) *Behavior characteristic* – the ability of a model to predict particular behavior (sudden peaks, long declines, etc.). Circumstances and pattern of behavior leading to a particular event are of greater interest than the event's exact timing.
Behavior prediction tests	A group of tests focusing on a model-predicted future behavior	*Pattern prediction* – a model should generate the qualitatively correct pattern of future behavior; *Event prediction* – focusing on a single event.
Behavior anomaly	Discovering sharply inconsistent behavior	Tracing anomalous behavior to elements of a model's structure. This test can also be used to defend a particular assumption by demonstrating anomalous behavior if an assumption is changed.
Family member	Testing model on similar cases	A model represents a family of social systems. If a model's parameters are changed appropriately, the characteristics of a different member of a family should be displayed. For example, Urban Dynamics should be able to portray the behavior of Berlin, Moscow, and other big cities.
Surprise behavior	Observing (often as a surprise) modeled behavior that is present, but unrecognized in the real system	Identification and understanding of previously unrecognized real system behavior displayed by a model.
Extreme policy	Evaluating model behavior under extreme values of rate equations (policies)	A model should behave in a way the real system would under the same extreme conditions
Boundary adequacy (behavior)	A boundary test conducted as a behavior test – evaluating additional structures that might influence model behavior	Analyzing model behavior with and without additional structure.
Behavior sensitivity	Analyzing the sensitivity of the model's behavior to changes in the parameters' values	Likely changes in parameter values should not cause a model to fail. The real system should be evaluated if it is likewise sensitive to the parameters in question.

XV.7. Design and testing of policies

Once a user has confidence in a model, it can be used for policy evaluation by testing the sensitivity of output to changes in the parameters. The model can also be used for policy design: the creation of entirely new strategies and structures. Because of the non-linear nature of systems, the sum of the effects of the individual policies might not be equal to the effect of the policies in combination (Sterman 2000).

Table XV.3 below contains several tests of policy implications (Forrester and Senge 1980).

Table XV.3: SD model tests of policy implications (based on Forrester and Senge 1980)

Test	**Purpose**	**Methodology**
System improvement	Identifying policies that lead to improvements in the real-life system	After one is confident in a model, policies that produce positive changes in it can be recommended for implementation in a real system. The problems with this test are proving that positive changes resulted from policy changes (were other conditions constant?); slow accumulation of test results in real systems (change occurs over extended period of time).
Changed behavior prediction	Analyzing whether a model correctly predicts a change in the system's behavior following the policy change	Two possible ways are: first, changing model policy and evaluating plausibility of model behavior; second, introducing model policy changes that occurred in a real system and comparing model's and observed behaviors.
Boundary adequacy (policy)	Analyzing if the added model structure results in changes in policy recommendations	If added structure does not produce significant changes in policy recommendations, the original boundary was adequate.
Policy sensitivity	Analyzing sensitivity of policy recommendations to the changes in parameter values	If a model recommends the same policy regardless of the parameters' values within a reasonable range, a model indicates the low risk of adopting it for policy making.

Appendix XVI: SD Model Equations

Lectures_Missed(t) = Lectures_Missed(t - dt) + (Absense_rate) * dt
INIT Lectures_Missed = 0

INFLOWS:
Absense_rate = Normal_absense_rate+Delta_SWC*B58+Delta_SWI*B59
Like_Subject(t) = Like_Subject(t - dt) + (Change_in_L_Subject) * dt
INIT Like_Subject = 3

INFLOWS:
Change_in_L_Subject = Delta_SWC*B10_8+Delta_SWI*B10_9
Perc_Wkload(t) = Perc_Wkload(t - dt) + (Change_in_P_Wkload) * dt
INIT Perc_Wkload = 0

INFLOWS:
Change_in_P_Wkload = Delta_HW_time*B74
Real_HW_time(t) = Real_HW_time(t - dt) + (Change_in_HW_rate) * dt
INIT Real_HW_time = 0

INFLOWS:
Change_in_HW_rate =
B46*Delta_Score+B4_10*Delta_Like_Subject+Lectures_Missed*B45
Relative_knowledge(t) = Relative_knowledge(t - dt) + (Knowl_gain_rate) * dt
INIT Relative_knowledge = 0

INFLOWS:
Knowl_gain_rate = B64*HW_time_diff
Satisf_W_Crse(t) = Satisf_W_Crse(t - dt) + (Change_in_SWC) * dt
INIT Satisf_W_Crse = 3

INFLOWS:
Change_in_SWC = Delta_Score*B86+Delta_SWI*B89+Delta_Workload*B87
Satisf_W_Instr(t) = Satisf_W_Instr(t - dt) + (Change_in_SWI) * dt
INIT Satisf_W_Instr = 3

INFLOWS:
Change_in_SWI = Delta_Workload*B97+Delta_Score*B96
Assignment_dfclty = 3
Assignment_rate = 1
Avg_Like_S = 3
Avg_SWC = 3
Avg_SWI = 3
B10_8 = 0.3577
B10_9 = 0.3332

B45 = -0.2000
B46 = -0.2374
B4_10 = 0.0437
B58 = -0.0027
B59 = -0.0951
B64 = 1.7578
B65 = -0.1382
B74 = 0.3346
B86 = 0.2417
B87 = -0.0902
B89 = 0.7279
B96 = 0.0831
B97 = -0.2759
Delta_HW_time = Real_HW_time-Normal_HW_Time
Delta_Like_Subject = Like_Subject-Avg_Like_S
Delta_Score = Test_score-Expected_score
Delta_SWC = Satisf_W_Crse-Avg_SWC
Delta_SWI = Satisf_W_Instr-Avg_SWI
Delta_Workload = Perc_Wkload-Normal_Workload
Expected_HW_time = Assignment_dfclty*Assignment_rate
Expected_score = 80
HW_time_diff = Real_HW_time-Expected_HW_time
Normal_absense_rate = 0.5
Normal_HW_Time = 3
Normal_Workload = 3
Test_score = max(min(Relative_knowledge+100-B65*Lectures_Missed,100),0)

Appendix XVII: Before and After Questions Appearing in MBKPs Administered in Winter 2005

1. Which of these statements about student performance evaluation in ENGG 401-B1 is TRUE?

   a. Quizzes are open book
   b. There are three non-cumulative exams in the course
   c. There are six quizzes in the course
   d. Late assignments are not accepted

2. The main topic that ENGG 401 covers is:

   a. Quality management in engineering practice
   b. Engineering economic analysis
   c. Bookkeeping for engineers
   d. Corporate social responsibility

3. Which of the following statements regarding the logistics of ENGG 401-B1 is FALSE?

   a. Course notes used last year are identical to the ones used this year
   b. If at all possible, appropriate material from the notes should be read before the lecture
   c. Access to and use of a spreadsheet program are required for efficient problem solving
   d. Assignment solutions and other files will be posted on the course web-page

4. What is the EMV of a project that results in a gain of \$10 million with 60% probability or in a loss of \$5 million with 40% probability?

   a. \$5 million
   b. \$4 million
   c. \$3 million
   d. 20%

5. Which of the following terms is used to denote a statement that we refer to in the class as the "balance sheet"?

a. Statement of operations
b. Consolidated statement of operations
c. Statement of financial position
d. Statement of earnings

6. Which of the following financial statements is given for a specific time interval?

a. Income statement
b. Statement of cash flow
c. Balance sheet
d. None of the above

7. Which of the following statements about depreciation is TRUE?

a. Too long a depreciation period understates income
b. If SV=0, D_1 using the SL method will be the same as D_1 using the DB method
c. CCA is equivalent to BV
d. Depreciation is a cash expense

8. Assuming book depreciation, if an asset purchased for $200,000 is depreciated over five years by using the declining balance method, the depreciation amount in the second year will be:

a. $160,000
b. $40,000
c. $32,000
d. $22,000

9. The shift from the DDB to the SL method is made:

a. In the second year of depreciation
b. In the third year of depreciation
c. When the amount depreciated with SL becomes smaller than the DB amount
d. When the amount depreciated with SL becomes larger than the DB amount

10. Which of the following statements about the balance sheet is TRUE?

a. Accrued wages payable to employees are recorded as assets
b. The balance sheet is accurate for at least a week at the time
c. A two-year line separates the current from long-term items
d. None of the above

11. Which of the following statements about the balance sheet is FALSE?

a. Cash is more liquid than the value of a patent
b. Debt is serviced before equity
c. Interest on the long term loan is recorded under current assets
d. Retained earnings are recorded under shareholder equity

12. You took out a 5-year loan of $10 million two years ago, and paid two annual payments according to the straight line schedule. What does the current balance sheet show under the long-term entry below the line?

a. $6 million
b. $5 million
c. $4 million
d. $2 million

13. At the end of which month does the "Magic Box" company discussed in class get into trouble with the bank over the short-term credit line?

a. First
b. Second
c. Third
d. Fourth

14. Which of the following statements is FALSE?

a. An increase in payables is a source of cash
b. A decrease in receivables is a source of cash
c. An increase in long-term debt is a use of cash
d. A decrease in accrued wages is a use of cash

15. The difference between sources and uses of funds over a certain period of time should equal:

a. Depreciation
b. Net income
c. Retained earnings
d. Zero

16. Which of the following statements is TRUE?

a. Horizontal analysis is performed with one financial statement only
b. Vertical analysis is performed year over year
c. Horizontal analysis is performed year over year
d. Vertical and horizontal analysis require stock market information only

17. The "quick" ratio is:

a. Always higher than the current ratio
b. Calculated by adding inventory to the current ratio
c. A profitability indicator
d. None of the above

18. Which of the following statements is TRUE?

a. The lower the DSO, the better
b. The lower the Inventory Turnover, the better
c. During downturns, Inventory Turnover increases
d. None of the above

19. Which of the following statements is TRUE regarding preferred shares?

a. They are riskier than common shares
b. Lenders treat them as debt
c. Common shareholders treat them as equity
d. None of the above

20. How much should I invest today if I expect to get $3000 three years from now at the annual rate of 3%?

a. $1,000
b. $2,745
c. $3,000
d. $3,278

21. Which of the following statements regarding time value of money is FALSE?

a. Converting a present value into a future value is called “compounding”
b. A dollar today is worth more than a dollar tomorrow
c. Formula $F=P(1+i)^N$ uses simple interest
d. “P” stands for “Present Value”

Appendix XVIII: Statistical Analysis of Survey Data, Winter 2005

Table XVIII.1: Goodness-of-fit test, midterm marks, section X, Winter 2005

Midterm mark	TRUE fr	true, %	reported fr	reported, %	expected fr
below 60	38	21.0%	22	18.8%	24.56
60-69	31	17.1%	18	15.4%	20.04
70-79	50	27.6%	37	31.6%	32.32
80-89	42	23.2%	25	21.4%	27.15
90-100	20	11.0%	15	12.8%	12.93
Total	181	100%	117	100%	117.00

Goodnes-of-fit	
d.f. for ch-sq	4
Ho	distributions are the same
H1	distributions are different
chi-square	1.655
chi-sq critical	9.488
P-value	0.799
conclusion	Ho not rejected

Table XVIII.2: Goodness-of-fit test, midterm marks, section Y, Winter 2005

Midterm mark	TRUE fr	true, %	reported fr	reported, %	expected fr
below 45	12	13.2%	5	10.4%	6.33
45-54	19	20.9%	9	18.8%	10.02
55-64	28	30.8%	14	29.2%	14.77
65-74	24	26.4%	18	37.5%	12.66
75 and above	8	8.8%	2	4.2%	4.22
Total	91	100%	48	100%	48

Goodnes-of-fit	
d.f. for ch-sq	4
Ho	distributions are the same
H1	distributions are different
chi-square	3.844
chi-sq critical	9.488
P-value	0.427
conclusion	Ho not rejected

Table XVIII.3: Comparison of survey results between Winter 2005 and 2003-2004

Workload

F-Test Two-Sample for Variances

	03-04	*W05*
Mean	3.23958	2.9454
Variance	0.58475	0.2957
Observations	384	165
df	383	164
F	1.97692	
P(F<=f) one-tail	5.2E-07	
F Critical one-tail	1.24991	

t-Test: Two-Sample Assuming Unequal Variances

	03-04	*W05*
Mean	3.23958	2.9454
Variance	0.58475	0.2957
Observations	384	165
Hypothesized Mean Difference	0	
df	429	
t Stat	5.1082	
P(T<=t) one-tail	2.5E-07	
t Critical one-tail	1.64841	
P(T<=t) two-tail	4.9E-07	
t Critical two-tail	1.96551	

Satisfaction with instructor

F-Test Two-Sample for Variances

	03-04	*W05*
Mean	3.5677	3.6909
Variance	1.4105	0.7758
Observations	384	165
df	383	164
F	1.8181	
P(F<=f) one-tail	8.2E-06	
F Critical one-tail	1.2499	

t-Test: Two-Sample Assuming Unequal Variances

	03-04	*W 05*
Mean	3.56771	3.6909
Variance	1.41055	0.7758
Observations	384	165
Hypothesized Mean Difference	0	
df	413	
t Stat	-1.3462	
P(T<=t) one-tail	0.08949	
t Critical one-tail	1.64855	
P(T<=t) two-tail	0.17897	
t Critical two-tail	1.96572	

Table XVIII.3: continued

Satisfaction with the course

F-Test Two-Sample for Variances

	03-04	*W05*
Mean	3.46354	3.33333
Variance	1.30937	1.07724
Observations	384	165
df	383	164
F	1.21549	
P(F<=f) one-tail	0.07498	
F Critical one-tail	1.24991	

t-Test: Two-Sample Assuming Equal Variances

	03-04	*W05*
Mean	3.46354	3.33333
Variance	1.30937	1.07724
Observations	384	165
Pooled Variance	1.23977	
Hypothesized Mean Difference	0	
df	547	
t Stat	1.25629	
P(T<=t) one-tail	0.10477	
t Critical one-tail	1.64765	
P(T<=t) two-tail	0.20955	
t Critical two-tail	1.96431	

Attitude toward the subject

F-Test Two-Sample for Variances

	03-04	*W05*
Mean	3.6589	3.36364
Variance	1.1392	1.00111
Observations	384	165
df	383	164
F	1.1379	
P(F<=f) one-tail	0.1706	
F Critical one-tail	1.2499	

t-Test: Two-Sample Assuming Equal Variances

	03-04	*W05*
Mean	3.6589	3.36364
Variance	1.1392	1.00111
Observations	384	165
Pooled Variance	1.0978	
Hypothesized Mean Difference	0	
df	547	
t Stat	3.0269	
P(T<=t) one-tail	0.0013	
t Critical one-tail	1.6476	
P(T<=t) two-tail	0.0026	
t Critical two-tail	1.9643	

Table XVIII.4: Comparison between sections X and Y, Winter 2005 semester

Workload

F-Test Two-Sample for Variances

	Y	X
Mean	2.64583	3.06838
Variance	0.27615	0.25391
Observations	48	117
df	47	116
F	1.08762	
P(F<=f) one-tail	0.35238	
F Critical one-tail	1.47021	

t-Test: Two-Sample Assuming Equal Variances

	Y	X
Mean	2.64583	3.06838
Variance	0.27615	0.25391
Observations	48	117
Pooled Variance	0.26032	
Hypothesized Mean Difference	0	
df	163	
t Stat	-4.83157	
P(T<=t) one-tail	1.6E-06	
t Critical one-tail	1.65425	
P(T<=t) two-tail	3.1E-06	
t Critical two-tail	1.97462	

Satisfaction with instructor

F-Test Two-Sample for Variances

	Y	X
Mean	3.20833	3.8889
Variance	0.97695	0.5613
Observations	48	117
df	47	116
F	1.72871	
P(F<=f) one-tail	0.00959	
F Critical one-tail	1.47021	

t-Test: Two-Sample Assuming Unequal Variances

	Y	X
Mean	3.20833	3.8889
Variance	0.97695	0.5613
Observations	48	117
Hypothesized Mean Difference	0	
df	70	
t Stat	-4.2885	
P(T<=t) one-tail	2.8E-05	
t Critical one-tail	1.66692	
P(T<=t) two-tail	5.7E-05	
t Critical two-tail	1.99444	

Table XVIII.4: continued

Satisfaction with the course

F-Test Two-Sample for Variances

	Y	X
Mean	3	3.47009
Variance	1.10638	1.00987
Observations	48	117
df	47	116
F	1.09557	
P(F<=f) one-tail	0.3413	
F Critical one-tail	1.47021	

t-Test: Two-Sample Assuming Equal Variances

	Y	X
Mean	3	3.47009
Variance	1.10638	1.00987
Observations	48	117
Pooled Variance	1.0377	
Hypothesized Mean Difference	0	
df	163	
t Stat	-2.6922	
P(T<=t) one-tail	0.00392	
t Critical one-tail	1.65425	
P(T<=t) two-tail	0.00784	
t Critical two-tail	1.97462	

Attitude toward the subject

F-Test Two-Sample for Variances

	Y	X
Mean	3.25	3.4126
Variance	1.0426	0.9841
Observations	48	117
df	47	116
F	1.058	
P(F<=f) one-tail	0.3954	
F Critical one-tail	1.4702	

t-Test: Two-Sample Assuming Equal Variances

	Y	X
Mean	3.25	3.4126
Variance	1.0426	0.9841
Observations	48	117
Pooled Variance	1.0019	
Hypothesized Mean Difference	0	
df	163	
t Stat	-0.9341	
P(T<=t) one-tail	0.1758	
t Critical one-tail	1.6543	
P(T<=t) two-tail	0.3517	
t Critical two-tail	1.9746	

Table XVIII.5: Comparison between instructor Y's results, Winter 2004 and Winter 2005

Workload

F-Test Two-Sample for Variances

	W04	*W05*
Mean	3.58333	2.64583
Variance	0.51836	0.27615
Observations	60	48
df	59	47
F	1.87708	
P(F<=f) one-tail	0.01347	
F Critical one-tail	1.59345	

t-Test: Two-Sample Assuming Unequal Variances

	W04	*W05*
Mean	3.58333	2.64583
Variance	0.51836	0.27615
Observations	60	48
Hypothesized Mean Difference	0	
df	105	
t Stat	7.81453	
P(T<=t) one-tail	2.2E-12	
t Critical one-tail	1.6595	
P(T<=t) two-tail	4.5E-12	
t Critical two-tail	1.98282	

Satisfaction with instructor

F-Test Two-Sample for Variances

	W04	*W05*
Mean	2.7666	3.2083
Variance	1.2327	0.9769
Observations	60	48
df	59	47
F	1.2618	
P(F<=f) one-tail	0.2054	
F Critical one-tail	1.5934	

t-Test: Two-Sample Assuming Equal Variances

	W04	*W05*
Mean	2.7666	3.2083
Variance	1.2327	0.9769
Observations	60	48
Pooled Variance	1.1193	
Hypothesized Mean Difference	0	
df	106	
t Stat	-2.155	
P(T<=t) one-tail	0.0166	
t Critical one-tail	1.6593	
P(T<=t) two-tail	0.0333	
t Critical two-tail	1.9826	

Table XVIII.5: continued

Satisfaction with the course

F-Test Two-Sample for Variances

	W04	W05
Mean	2.71667	3
Variance	1.22345	1.1063
Observations	60	48
df	59	47
F	1.10581	
P(F<=f) one-tail	0.36287	
F Critical one-tail	1.59345	

t-Test: Two-Sample Assuming Equal Variances

	W04	W05
Mean	2.71667	3
Variance	1.22345	1.1063
Observations	60	48
Pooled Variance	1.17154	
Hypothesized Mean Difference	0	
df	106	
t Stat	-1.3518	
P(T<=t) one-tail	0.08966	
t Critical one-tail	1.65935	
P(T<=t) two-tail	0.17933	
t Critical two-tail	1.9826	

Attitude toward the subject

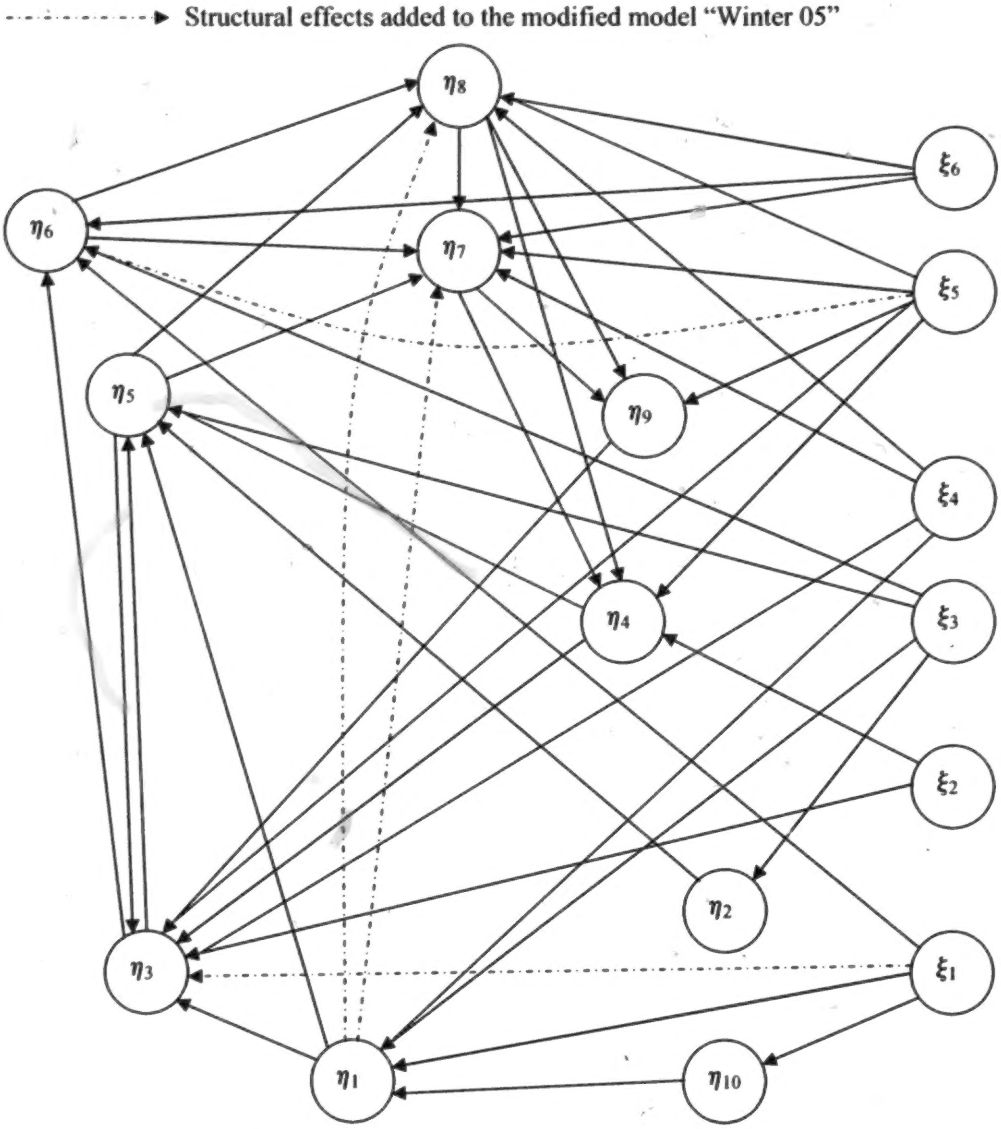
F-Test Two-Sample for Variances

	W04	W05
Mean	3.3	3.25
Variance	1.2644	1.0425
Observations	60	48
df	59	47
F	1.2128	
P(F<=f) one-tail	0.2479	
F Critical one-tail	1.5934	

t-Test: Two-Sample Assuming Equal Variances

	W04	W05
Mean	3.3	3.25
Variance	1.2644	1.0425
Observations	60	48
Pooled Variance	1.166	
Hypothesized Mean Difference	0	
df	106	
t Stat	0.2391	
P(T<=t) one-tail	0.4057	
t Critical one-tail	1.6594	
P(T<=t) two-tail	0.8115	
t Critical two-tail	1.9826	

Appendix XIX: SEM Model "Winter 05" in a Path-Diagram Form



Structural effects added to the modified model "Winter 05"

Variable	Label	Variable	Label
ξ_1	Importance of having fun while in university	η_1	Extra-curricular activities
ξ_2	Importance of succeeding academically	η_2	Language practice
ξ_3	Language background	η_3	Time devoted to self-studying
ξ_4	Age	η_4	Attendance
ξ_5	Instructor's teaching experience	η_5	Student knowledge
ξ_6	Academic background in the discipline	η_6	Perceived course workload
		η_7	Satisfaction with the course in general
		η_8	Satisfaction with the instructor
		η_9	Attitude toward the subject
		η_{10}	Balancing acad. perf. and "fun"

Appendix XX: *S* matrix, Winter 2005

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
EXTRACTV	0.9563									
SPEKENGL	0.0747	0.7829								
READTEXT	-0.0932	0.0004	0.4035							
HWRKTIME	-0.0952	-0.0026	0.0819	0.2118						
LTRSMISS	0.0213	0.0398	-0.0431	-0.0627	0.3564					
MIDTSCRE	0.0549	0.0793	0.0688	0.0238	-0.1200	1.4903				
WORKLOAD	0.0113	-0.0322	0.0699	0.0051	-0.0063	-0.0377	0.2940			
SATISFCT	0.2303	-0.0646	0.0667	0.0424	-0.0505	0.5535	0.0303	1.0707		
INTRGJOB	0.1269	0.0338	-0.0688	0.0126	0.0432	0.2046	0.0074	0.4485	0.7711	
LKESBJCT	0.0860	-0.0501	0.0413	0.0066	0.0055	0.4209	-0.0590	0.6121	0.3791	0.9950
DOWLHVFN	0.1321	0.0089	0.0085	-0.0085	0.0008	-0.0032	-0.0012	0.0263	-0.0170	0.0061
HAVEFUNU	0.0088	-0.0026	-0.0060	-0.0154	0.0028	0.0002	-0.0053	0.0070	0.0089	0.0073
DOWELLUN	-0.0068	0.0102	0.0214	0.0051	-0.0123	0.0411	0.0031	0.0242	-0.0047	-0.0105
LANGBGND	-0.1100	-0.0795	0.0469	0.0032	-0.0181	-0.0442	0.0520	0.0263	0.0058	0.0088
STUDTAGE	-0.0339	-0.0319	0.0679	0.0291	0.0186	-0.0683	0.0388	-0.0202	0.0218	-0.0485
INSTEXPR	-0.0021	-0.0363	-0.0355	0.0068	0.0210	-0.0252	0.0872	0.0970	0.1404	0.0331
SMLRCRSE	-0.1133	-0.1028	0.0199	-0.0106	0.0118	0.0223	0.0410	0.0848	0.0141	0.0298

	Y11	X1	X2	X3	X4	X5	X6
EXTRACTV							
SPEKENGL							
READTEXT							
HWRKTIME							
LTRSMISS							
MIDTSCRE							
WORKLOAD							
SATISFCT							
INTRGJOB							
LKESBJCT							
DOWLHVFN	0.1156						
HAVEFUNU	-0.0007	0.0179					
DOWELLUN	0.0109	-0.0002	0.0516				
LANGBGND	-0.0129	0.0008	-0.0025	0.2176			
STUDTAGE	-0.0028	-0.0004	0.0085	0.0420	0.2440		
INSTEXPR	-0.0145	-0.0013	-0.0098	0.0208	-0.0012	0.2063	
SMLRCRSE	-0.0230	0.0021	-0.0075	0.0780	0.0400	0.0247	0.5100

Appendix XXI: LISREL Syntax, Model “Winter 05”

```
Title  Kosta G.  Thesis model based on ENGG 401 Winter 2005 surveys
DA NI=17 NO=165 MA=CM
CM sy

[covariance matrix, Appendix Y1, goes here]

LA
!indicators' labels
EXTRACTV SPEKENGL READTEXT HWRKTIME LTRSMISS
MIDTSCRE WORKLOAD SATISFCT INTRGJOB LIKESBJCT DOWLHVFN
HAVEFUNU DOWELLUN LANGBGND STUDTAGE INSTEXPR SMLRCRSE
MO NY=11 NX=6 NE=10 NK=6 LY=FU,FI LX=ID BE=FU,FI GA=FU,FI PH=FU,FR
PS=DI,FR TE=DI,FI TD=DI,FI
!Model specs
! NY-number of Y indicators, NX-numb of X indicators, NE-num of eta variables, NK-num
of ksi vars

! Lambda X is identity matrices since there is a single indicator per concept
! Lambda Y is full fixed (at 0), and values 1 will be assigned to single indicators,
! and one of multiple indicators for concept ETA 3 ("Time to course") will be freed
! Beta and Gamma are full, and fixed (at initial values 0). Later the coefficients to be
!estimated will be freed

!PSI is the matrix of covariances among the errors of exogenous vars ETA. It is diagonal to
exclude
!covariances between ETAs' errors (as they are assumed to be independent). The matrix is
free, so that LISREL would estimate error variances

! Tay E and Tay D are the matrices of covariances between the errors of indicators. These
matrices are
!diagonal to exclude possible covariances between indicators' errors.

FR BE(1,10), BE(3,1), BE(3,4), BE(3,5), BE(3,9), BE(4,7), BE(4,8), BE(5,1)
FR BE(5,2), BE(5,3), BE(5,4), BE(6,3), BE(7,5), BE(7,6), BE(7,8)
FR BE(8,5), BE(8,6), BE(9,7), BE(9,8)
! free up beta elements that will be estimated by LISREL


FR GA(1,1), GA(1,3), GA(1,4), GA(2,3), GA(3,2), GA(3,4), GA(3,5)
FR GA(4,2), GA(4,5), GA(5,3), GA(6,1), GA(6,3), GA(6,6), GA(7,4)
FR GA(7,5), GA(7,6), GA(8,4), GA(8,5), GA(8,6), GA(9,5), GA(10,1)

FR GA(3,1)
FR BE(7,1)
FR BE(8,1)
FR GA(6,5)
!elements freed in the modified model
```

```
VA 1.0 LY(1,1), LY(2,2), LY(4,3), LY(5,4), LY(6,5), LY(7,6), LY(8,7), LY(9,8)
VA 1.0 LY(10,9), LY(11,10)

FR LY(3,3)
!Multiple (second) indicator of the concept "Time to course"


VA 0.10560 TE(1,1)
! 10% error var to the indicator "extra activities"
VA 0.16692 TE(2,2)
! 20% to Speak English
FR TE(3,3)
! free indicator, Read text
VA 0.15127 TE(4,4)
! 50%, multiple indicator, HW time
VA 0.08244 TE(5,5)
! 20% to Lectures missed
VA 0.14492 TE(6,6)
! 10% to Test Score
VA 0.03218 TE(7,7)
! 10% to Workload
VA 0.22914 TE(8,8)
! 20% to Satisfaction with course
VA 0.17497 TE(9,9)
! 20% to Instructor did a good job
VA 0.20652 TE(10,10)
! 20% to Like subject
VA 0.01193 TE(11,11)
! 10% to Do well and have fun


VA 0.00026 TD(1,1)
! 5% to Have fun university
VA 0.00220 TD(2,2)
! 5% to Do well in university
VA 0.01987 TD(3,3)
! 10% to Language background
VA 0.02410 TD(4,4)
! 10% to Age
VA 0.03711 TD(5,5)
! 20% Instructor's experience
VA 0.02496 TD(6,6)
!5% to Took similar course

OU ME=ML ALL ND=4
!output: method of estimation-max likelihood, numb of digits-4
```

Appendix XXII: Data Collected From MBKPs, Winter 2005

Question number	Sample size	Number of IBCA answers	Number of IA answers
1	102	49	2
2	102	42	11
3	102	47	16
4	113	88	5
5	113	52	25
6	113	29	60
7	88	33	9
8	88	49	10
9	88	56	5
10	88	55	14
11	88	36	27
12	88	42	47
13	68	57	6
14	68	41	17
15	68	44	6
16	88	64	6
17	88	69	11
18	88	71	7
19	78	42	7
20	78	20	7
21	78	54	7